Mage-VL: An Efficient Codec-Native Streaming Multimodal Foundation Model

Microsoft Mage Team

Standard Vision-Language Models (VLMs) suffer from Moravec's paradox: they excel at complex offline visual reasoning, but fail and suffer computational inefficiency on simple streaming perception tasks. We present Mage-VL, an efficient codec-native streaming foundation model for real-time multimodal understanding and interaction. At its core, our custom tokenizer, Mage-ViT, replaces uniform frame sampling by selectively encoding dynamic, entropy-rich regions using motion vectors and residual energy across sparse anchor (I) and predicted (P) frames. Operating at a 16×16 patch level, this reduces visual token consumption by over 75% while preserving spatio-temporal context. Trained from scratch on approximately 560M unlabeled images and 100M unlabeled video frames, Mage-ViT matches or outperforms flagship encoders trained on billions of image-text pairs. We establish AI4AI data pipelines, encompassing prompt-code joint optimization for multimodal captioning and AI-driven performance diagnosis to guide training recipes. Furthermore, through a bio-inspired dual-system architecture—a lightweight System 1 event gate and a causal System 2 decoder—Mage-VL enables proactive streaming perception. Extensive evaluations show that Mage-VL-4B matches Qwen3-VL-4B on static tasks while achieving strong gains in video understanding and 2D/3D spatial reasoning with up to $3.5\times$ wall-clock inference speedup, comprehensively surpassing the 15B Phi-4-reasoning-vision baseline. Beyond model artifacts, we deliver seven key empirical findings covering pre-training data efficiency, variable-resolution scaling, codec system acceleration, VideoQA SFT redundancy, motion-spatial synergy, AI4AI data pipelines, and Zero-Vision SFT for multimodal RL.

Project Page: https://microsoft.github.io/Mage
Code: https://github.com/microsoft/Mage
Model: https://huggingface.co/collections/microsoft/mage
Date: July, 2026

1 Introduction

Recent vision-language models (VLMs) have substantially advanced multimodal intelligence [1–3]. These models can decipher complex code [1], read dense documents [2], and solve geometry problems [3] with high proficiency. However, extending these capabilities to continuous video streams exposes a modern manifestation of **Moravec's paradox** [4]. Existing open-source VLMs—such as Qwen-VL [5, 6], InternVL [7], LLaVA-OneVision [8], Keye-VL [9], and Cambrian [10]—perform well on complex reasoning over sparsely sampled keyframes, yet remain slow and computationally heavy when responding to dynamic visual events.

This paradox stems from a structural mismatch between continuous visual input and current model architectures. Physical perception is continuous and event-driven, whereas mainstream VLMs rely on vision encoders pre-trained on web-scale static images. Kimi-VL [11], for instance, initializes its vision encoder from SigLIP2 [12], which is trained on over 10 billion image–text pairs. When processing videos, these models typically use uniform frame sampling at fixed resolutions. This strategy ignores spatio-temporal redundancy: static background regions are

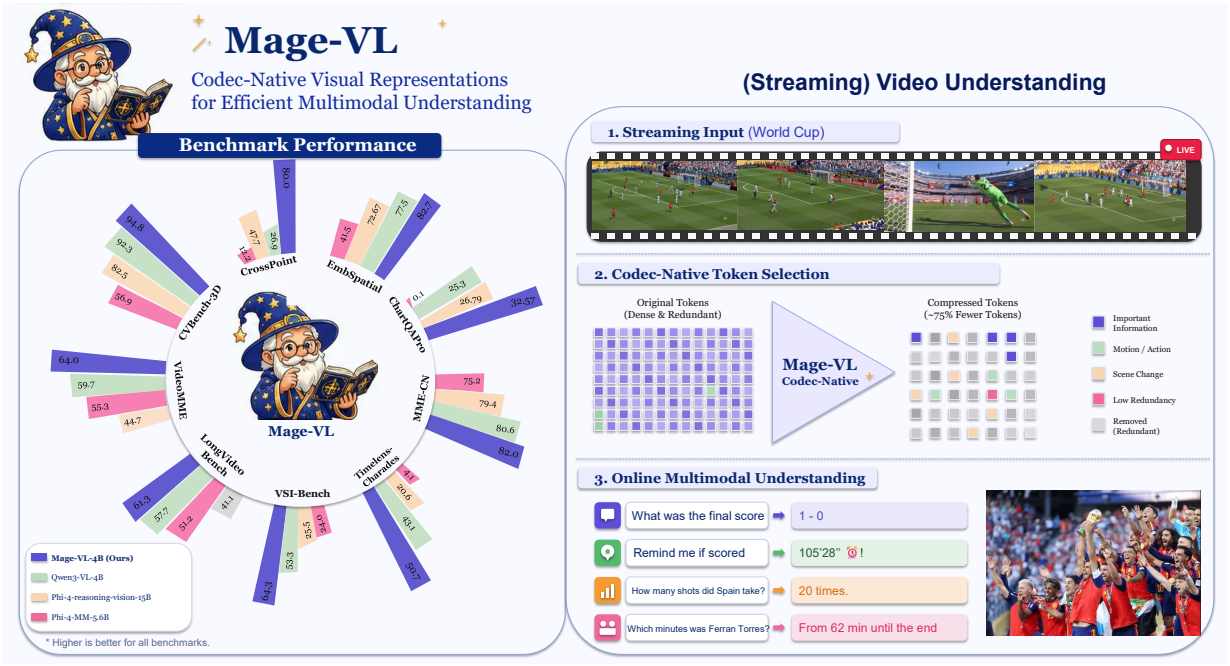

Figure 1 **Overview of Mage-VL. (Left)** Performance comparison of Mage-VL-4B against Qwen3-VL-4B, Phi-4-MM (5.6B), and Phi-4-reasoning-vision (15B) across spatial, video, and visual understanding benchmarks. **(Right)** Codec-native video processing workflow: given continuous streaming video inputs, Mage-VL uses motion and residual signals to select dynamic visual patches, reducing visual tokens by approximately 75% while supporting real-time online video question answering.

repeatedly encoded even when only a small portion of the scene changes between frames. As a result, computation scales rapidly with video duration.

Biological visual processing suggests a more efficient approach. Natural vision operates under strict resource constraints by filtering out redundant background signals at the sensory front-end and prioritizing dynamic motion [13–15]. Higher-level perception further relies on a dual-system process: a fast, low-latency pathway that determines when to react [16], combined with a deliberate reasoning process for complex tasks [17].

Inspired by this efficient dual-process design, we present **Mage-VL**, a codec-native streaming foundation model. Instead of uniform frame sampling, Mage-VL models video streams by adapting principles from classical video codecs. Built upon the OneVision Encoder architecture [18], it uses temporal motion vectors and residual energy to selectively encode dynamic regions rich in visual information. Incoming video is divided into anchor frames (I-frames), which are fully processed, and predicted frames (P-frames), which retain only the patches affected by motion. Operating at a 16×16 patch level, this selective encoding significantly cuts visual token count while preserving temporal context.

Unlike standard VLMs that inherit vision encoders pre-trained on billions of static images, we train **Mage-ViT** from scratch using approximately 560 million unlabeled images and 100 million unlabeled video frames. Despite this small training volume, Mage-ViT performs on par with, and on several image and video benchmarks better than, established encoders trained on billions of image–text pairs (such as SigLIP2 [12] and MoonViT [11]). This finding demonstrates that alignment with temporal video structures and data efficiency can matter more than sheer pre-training dataset scale. Furthermore, while Mage-ViT is trained using H.265, it exhibits strong robustness across codec families: when paired with neural codecs such as DCVC [19] at inference time, it maintains comparable performance while further reducing visual token overhead.

Building on Mage-ViT, we pre-train and fine-tune the Mage-VL foundation model through a progressive five-stage curriculum. While the earlier four stages inherit structural archetypes from

LLaVA-OneVision2 [20], they differ fundamentally in data composition and strategic objectives. In **Stage 1**, we warm up the model with dense image captions and short-video captions using roughly four times the captioning data of LLaVA-OneVision2, endowing the language model with broad visual understanding capabilities. In **Stage 2**, we introduce image instruction data, temporal grounding tasks, spatial supervision, and additional video captions to establish instruction following and grounding capabilities. In **Stage 3**, we train on long-video captions to extend the model's perceptual context window. This teaches the model about event ordering and long-range dependencies. In **Stage 4**, we switch video supervision to codec video streams. Benefiting from our bio-inspired predictive patch mechanism, codec tokenization consumes approximately 1/8 or fewer visual tokens compared to dense frame sampling, allowing us to train on video sequences that are eight times longer than those in Stage 3. In **Stage 5**, we perform streaming training: we convert the earlier video captions into a streaming format and train a proactive gate on top of the Mage-ViT to serve as System 1. This gate decides when the model should respond to incoming streaming inputs. Once triggered, the full model acts as System 2 to perform comprehensive reasoning and generate the response. The resulting Mage-VL supports continuous perception, event-triggered interaction, and conventional user-initiated question answering within a single model. Notably, Mage-VL achieves strong performance within the 4B parameter class, particularly on video understanding and spatial reasoning. It also substantially outperforms larger Phi-family baselines, including Phi-4-multimodal-instruct (5.6B) [21] and Phi-4-reasoning-vision (15B) [22], across most evaluated dimensions, despite using considerably fewer parameters.

Beyond the model itself, this work provides systematic empirical studies and practical insights into efficient multimodal modeling, summarized in seven key findings: First, we show that web-scale pre-training is not essential for visual encoders; a custom backbone trained on merely 560M unlabeled images and 100M video frames can achieve performance competitive with encoders trained on multi-billion-scale image-text corpora [12]. Second, variable-resolution pre-training enables continuous, monotonic performance scaling with visual token budgets without resolution degradation. Third, codec-native tokenization establishes a superior spatio-temporal accuracy-efficiency frontier, achieving up to 3.5× wall-clock inference speedups over uniform frame sampling. Fourth, we show that explicit long VideoQA SFT is redundant; fine-tuning solely on dense video detailed captions alongside short video SFT is fully sufficient for strong zero-shot long VideoQA capabilities. Fifth, dynamic video training significantly enhances static 2D/3D spatial reasoning, validating the principle that motion and spatial structure are inherently synergistic [23]. Sixth, we demonstrate an AI4AI data pipeline where agentic closed-loop feedback and prompt-code co-design systematically boost caption quality and downstream benchmark performance, inspiring SkillOpt-Lite [24]. Seventh, we reveal that bypassing visual SFT in favor of pure-text reasoning SFT unlocks potent multimodal RL capabilities, offering a compute-efficient path to narrow the agentic capability gap.

Overall, our main contributions are as follows:

- **Codec-Native Proactive Streaming Architecture:** We introduce Mage-VL, a unified streaming VLM that operates on continuous video via sparse codec-native updates. By encoding motion-salient patches across anchor and predicted frames, our architecture reduces visual token consumption to 1/8 or less compared to dense frame sampling. Building on this dynamic representation, a lightweight System 1 gate monitors incoming visual streams to trigger a causal System 2 LLM decoder for low-latency, event-driven interaction.

- **Highly Efficient Foundation Models & Open Release:** We present and open-source Mage-ViT and Mage-VL-4B. Notably, Mage-ViT is pre-trained from scratch on approximately 560M unlabeled images and 100M unlabeled video frames, yet achieves competitive performance compared to vision encoders pre-trained on billions of image-text pairs (e.g., SigLIP2). Driven by this efficient vision foundation, Mage-VL-4B achieves up to 3.5× wall-clock

inference speedups, matches Qwen3-VL-4B on static tasks while outperforming it on video and spatial reasoning.

- **AI4AI and Empirical Insights:** We provide seven foundational empirical insights covering pretraining data efficiency, resolution scaling laws, codec-driven system acceleration, VideoQA SFT redundancy via dense captions, motion-spatial synergy, AI4AI closed-loop data pipeline optimization, and Zero-Vision SFT for multimodal RL. These findings challenge conventional multimodal scaling practices and provide practical recipes for efficient foundation model development.

2 Related Work

2.1 Visual Tokenization: From Image ViTs to Codec-ViT

Most VLM visual encoders inherit the dense patch-grid representation introduced by Vision Transformers [25], including CLIP [26], SigLIP and SigLIP2 [12, 27], DINOv2 [28], and the visual encoders of recent Qwen-VL models [5, 6]. To reduce redundant tokens, prior work has explored adaptive pruning and merging, including DynamicViT [29], AdaViT [30], ToMe [31], FastV [32], and LLaVA-PruMerge [33].

For video, TimeSformer [34], ViViT [35], Video Swin Transformer [36], and VideoMAE [37] extend visual modeling into the temporal dimension. Video VLMs further reduce token cost through temporal compression or adaptive selection, including LLaMA-VID [38], Chat-UniVi [39], SlowFast-LLaVA [40], MovieChat [41], LongVU [42], and VideoChat-Flash [43]. However, these methods generally compress features after dense frame encoding or select entire frames, leaving substantial spatio-temporal redundancy unresolved.

Codec-based representations provide a more direct way to exploit temporal redundancy. CoViAR [44] demonstrated compressed-domain video recognition using motion vectors and residuals, while Video-LaVIT [45] introduced codec-derived motion information into multimodal modeling. OneVision Encoder [18] further develops codec-aligned patch sparsity, and LLaVA-OneVision-2 [20] incorporates video-native representations into VLM training.

Mage-ViT builds on this direction by representing videos with anchor frames and sparse predicted frames, allocating tokens only to temporally informative regions before expensive visual encoding. Unlike image-centric encoders, Mage-ViT is trained from scratch on both image and video data, while its codec-native design supports efficient long-video processing and incremental streaming.

2.2 Open Vision–Language Models

Modern VLMs typically connect a pretrained vision encoder to a decoder-only LLM through a lightweight projector. Early works such as Flamingo [46] and BLIP-2 [47] explored efficient vision–language alignment, while LLaVA [48], MiniGPT-4 [49], InstructBLIP [50], and LLaVA-1.5 [51] established visual instruction tuning as a standard paradigm.

Recent open VLMs scale this recipe with stronger visual encoders, larger multimodal corpora, and multi-stage training. Representative models include Qwen-VL [5, 6, 52], InternVL [7, 53], DeepSeek-VL [54], Pixtral [55], Molmo [56], Kimi-VL [57], LLaVA-OneVision [8, 20], and Keye-VL [9]. Compact models such as Phi-3.5-Vision [58] and Phi-4-Multimodal [21] provide strong smaller-scale baselines, while BAGEL [59], Emu3.5 [60], Ovis-U1 [61] further unify visual understanding and generation [62].

Video VLMs largely extend this image-centric paradigm by aggregating features from sampled frames. VideoChat [63], Video-ChatGPT [64], Video-LLaMA [65], and Video-LLaVA [66] introduced early video extensions, while LLaVA-OneVision [67] and LongVA [68] improve unified image–video modeling and long-context understanding.

Despite these advances, most video VLMs still rely on sampled frames followed by dense

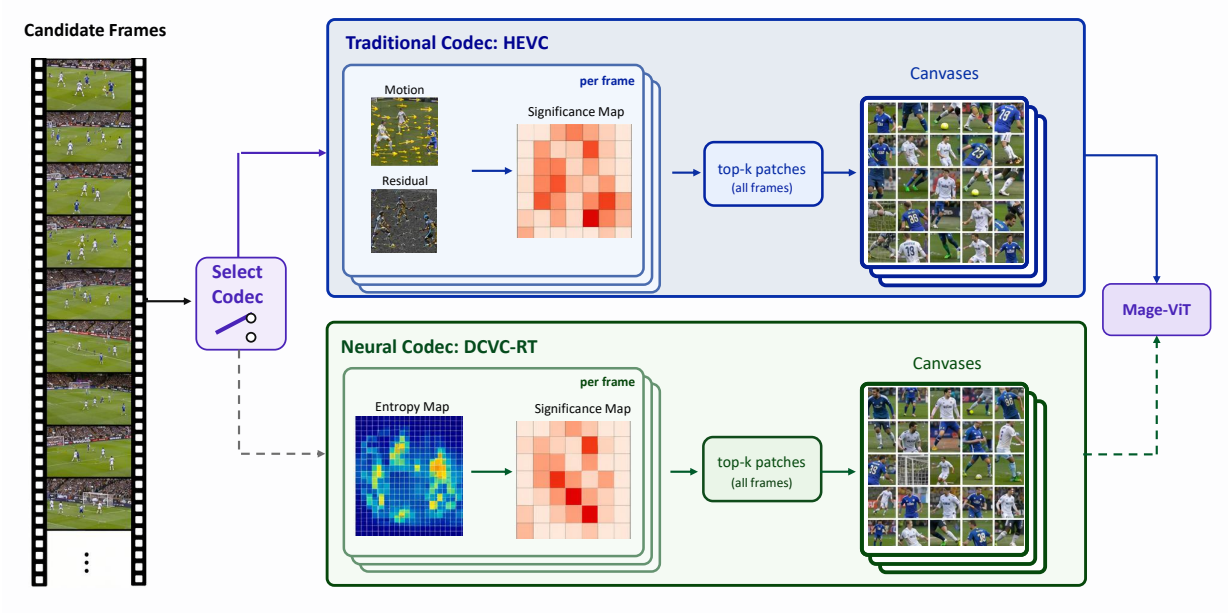

Figure 2 Codec-driven patchifier of Mage-ViT. Mage-ViT robustly supports both traditional codec HEVC/H.265 [81] and neural codec DCVC-RT [82] to select patches from video frames according to information density. The selected patches are composed into canvas, which is fed to Mage-ViT with corresponding spatial-temporal positional encoding information.

per-frame tokenization. Mage-VL instead trains a video-native visual encoder from scratch and directly consumes sparse codec-aligned visual streams, reducing temporal redundancy before it reaches the language model.

2.3 Streaming and Online Video-Language Models

Most video VLMs operate offline, processing sampled frames as a static multi-image input before generating a response. Recent work instead explores continuous video interaction and proactive response timing. VideoLLM-online [69] interleaves streaming perception and language generation, MMDuet [70] supports responses at arbitrary moments, and Dispider [71] decouples perception, decision, and reaction. StreamMind [72] further introduces event-gated response triggering, while JoyAI [73] enables real-time interaction through a multi-agent pipeline.

Another line of work focuses on memory and computational efficiency for long-running streams. Flash-VStream [74], StreamingVLM [75], and InternLM-XComposer2.5-OmniLive [76] maintain compressed or rolling visual states for long-term streaming understanding. Streaming supervision and evaluation have also been studied using temporally aligned data, such as LiveCC [77] and MatchTime [78], together with benchmarks including StreamingBench [79] and OVO-Bench [80].

Unlike prior methods that mainly add streaming mechanisms to conventional frame-based encoders, Mage-VL makes streaming native to the visual front-end. It processes rolling codec-token windows and integrates proactive response triggering with language generation through a dual-system architecture.

3 Mage-ViT

This section introduces Mage-ViT, the visual encoder at the core of Mage-VL. Mage-ViT is designed primarily for video, while single images are treated as a degenerate one-frame case. It follows the *codec-aligned sparsity* principle of OneVision-Encoder [18]: visual tokens should be allocated according to where the codec spends bits, since codec bit allocation provides a natural proxy for spatio-temporal importance. Section 3.1 describes the architecture, and Section 3.2

describes the pre-training data, objective and the two-stage recipe.

3.1 Architecture

Mage-ViT adopts a Codec-ViT architecture consisting of a codec-driven patchifier, a Vision Transformer trunk, and a shared 3D rotary positional encoding.

Codec-driven patchifier. Mage-ViT first divides the input into 16×16 pixel patches and then performs codec-guided sparse selection. For multi-frame inputs, a per-patch importance tensor $S \in \mathbb{R}_{\geq 0}^{T \times H \times W}$ is estimated to represent how many bits or how much information the codec assigns to that spatio-temporal patch. We adopt the *traditional codecs* HEVC/H.265 [81] as the default codec in patch selection, where S is defined as a weighted combination of the magnitude of motion vectors and the residual energy of P-frames. We note that our ViT can robustly support more diverse codecs such as the *neural codecs* DCVC-RT [82]. In this case, S is obtained directly from the codec's learned probability model, where the negative log-likelihood of each patch corresponds to the estimated number of bits required for coding. We observe that the resulting bit-allocation map provides a reliable proxy for local motion and temporal variation. The codec-driven patchifier is illustrated in Fig. 2.

Based on S, the patchifier keeps all I-frame patches, and selects the top-k P-frame patches by importance S within a token budget B. For a 64-frame clip with 16×16 tokens in each frame, we select $B = 4096$ tokens, corresponding to roughly 75% token reduction. We additionally support two other video input modes: *Chunk-wise patchification,* which follows conventional frame-centric paradigm to partition the input into C temporal chunks and samples one frame per chunk; *Collage patchification,* where one frame is first sampled from each temporal chunk and the sampled frames are vertically concatenated into a single tall 2D canvas.

ViT trunk. Mage-ViT uses a 24-layer pre-norm Vision Transformer with hidden dimension 1024, 16 attention heads, and GELU MLPs at 4× expansion ratio, implemented with Flash Attention 2 [83] for both training and inference. The trunk processes variable-length visual token sequences produced by codec-driven sparse selection. Its outputs are passed through the multimodal projector and subsequently consumed by the language model in Mage-VL (Section 4). During batched training, padding tokens are masked so that attention respects the valid sequence boundary of each sample.

Shared Positional encoding. Position information for the ViT trunk is provided by a shared 3D rotary positional encoding [84] over the un-pruned spatio-temporal grid. In this case, the spatial-temporal relations across patches are retained even when large fractions of the grid are dropped by the codec front-end.

3.2 ViT Pre-training

Mage-ViT is trained from scratch using a two-stage pre-training pipeline.

Data. We filter the image and video dataset following the filter rules in OneVision Encoder [18]. For images, we select source images from LAION-400M [85], COYO-700M [86], OBELICS [87], Zero250M [88], and ImageNet-21K [89]. For videos, we use videos from HowTo100M [90] and Panda-70M [91]. It contains a diverse set of publicly available images and videos, spanning natural photographs, artworks, documents, charts, indoor scenes, web videos, robotic trajectories, instructional content, and high-motion clips.

Objective. We train Mage-ViT with a large-scale *cluster-discrimination* objective over the visual representation space. The objective encourages semantically related samples to align with shared

visual concept prototypes. Specifically, we extract MetaCLIP features from both image and short-video data, perform K-means clustering over the combined corpus, and assign each training sample to its nearest cluster centers. Mage-ViT is then optimized using negative-sampling cluster discrimination against these visual prototypes. This objective encourages semantically related content, such as text regions, human faces, or traffic signs, to occupy nearby regions in the learned representation space.

Stage 1: Variable-resolution image pre-training. We first pre-train Mage-ViT on image data with resolutions ranging from 224 to 448 and aspect ratios from 1:2 to 2:1. Images of the same resolution are packed together to better utilize the target token budget, which we find improves optimization efficiency. This stage uses single-image spatial patchification. Since each sample contains only one frame, the temporal prediction pathway remains inactive. We use AdamW with an initial learning rate of 10^{-3}, cosine decay with warm-up, gradient clipping at 1.0, bf16 mixed precision, and a large effective batch size obtained through gradient accumulation. The negative-sampling ratio for cluster discrimination is set to $r = 0.1$.

Stage 2: Joint image and video pre-training. Starting from the Stage-1 checkpoint, we jointly train Mage-ViT on image and video data. Image resolution varies from 224 to 448, while videos are trained at resolution 256 with 64 frames per clip and a token budget of 4096. The three patchification modes are mixed during training with approximate proportions of 50% codec, 40% chunk-wise, and 10% collage. The learning rate is reduced to 5×10^{-5}. This stage activates codec-driven temporal sparsification and jointly optimizes spatial representation learning and spatio-temporal token selection. Training is conducted in bf16 mixed precision.

4 Mage-VL

Building upon the codec-native visual encoder introduced in Section 3, we develop Mage-VL, a unified multimodal model for image understanding, offline video reasoning, and proactive interaction over continuous video streams. The same model supports static images, videos of varying durations, event-triggered commentary, and conventional user-initiated question answering.

Section 4.1 presents the unified architecture across image, offline-video, and streaming modes. Section 4.2 describes the image, video, and streaming supervision, while Section 4.3 details the progressive five-stage training curriculum.

4.1 Unified Architecture

Codec-native visual encoder. Mage-VL uses Mage-ViT as its shared visual encoder across image, video, and streaming inputs. A still image is represented as a single spatial token sequence. For video, Mage-ViT produces temporally ordered codec-token windows composed of densely encoded anchor-frame patches and sparsely selected predicted-frame patches. Anchor frames preserve complete scene context, whereas predicted frames retain only patches associated with motion or visual changes. This codec-native representation avoids repeated computation over temporally static regions while preserving fine-grained spatial and temporal information. The resulting visual features are passed to the vision-language projector and causal language decoder.

Vision-language projector. A lightweight two-layer MLP maps Mage-ViT features into the token-embedding space of the language model. Because Mage-ViT applies shared 3D rotary positional encoding over the original spatio-temporal grid, retained tokens preserve their original spatial and temporal coordinates after codec-driven sparse selection.

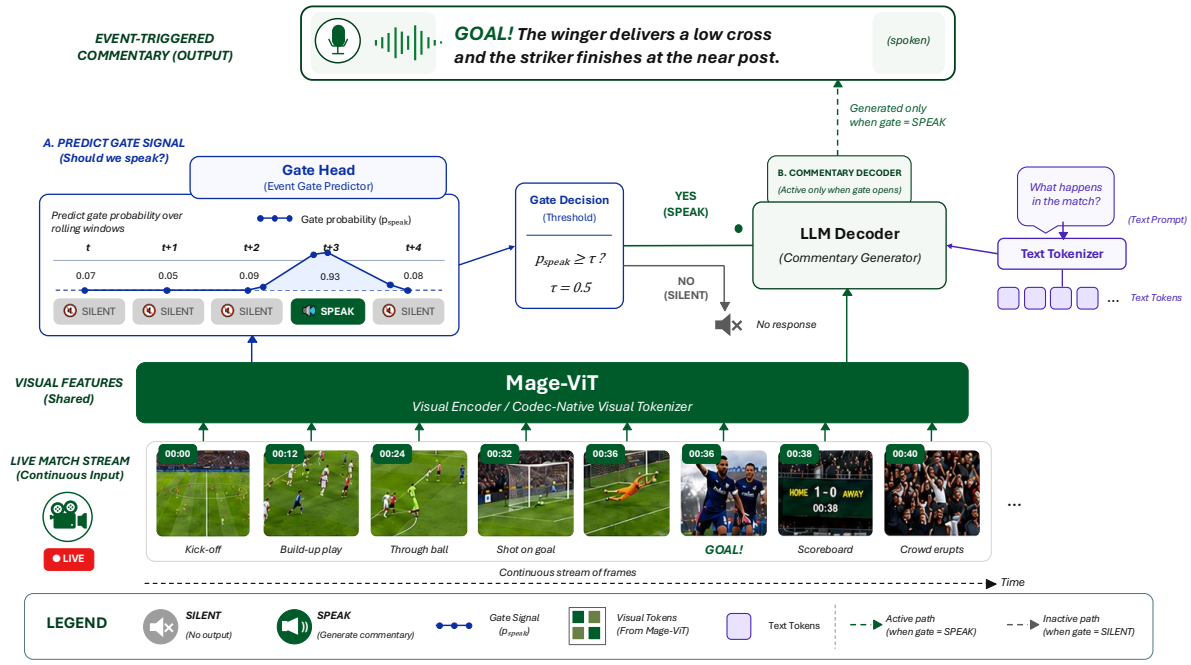

Figure 3 **Proactive streaming framework of Mage-VL.** Mage-ViT incrementally encodes a continuous video stream into codec-native visual features shared by the event gate and causal language decoder. The gate evaluates each rolling visual window and predicts whether the current prefix contains a response-worthy event. The model remains silent when the gate is closed; when it opens, the decoder generates an event-conditioned response from the recent visual context and the given text prompt.

Causal language decoder. The language backbone is initialized from Qwen3-4B-Instruct-2507 [92]. Projected visual tokens and text tokens are processed by the same causal decoder. For images, visual tokens form a single block preceding the text instruction. For videos, codec-token windows are concatenated in temporal order. This unified interface allows Mage-VL to process images, short videos, long videos, and ultra-long videos without modifying the language backbone or introducing modality-specific decoders.

Proactive streaming mechanism. As illustrated in Fig. 3, a lightweight cognition gate continuously evaluates the visual features of incoming codec windows. Given the current streaming representation $\mathbf{h}_t$, the gate predicts $p_{\text{speak}} = g(\mathbf{h}_t)$ and triggers generation when $p_{\text{speak}} \geq \tau$, where τ is a fixed inference threshold. Background segments and non-response-worthy moments keep the gate closed, whereas relevant events activate language generation.

During streaming inference, newly arrived codec windows are incrementally processed by the perception pathway. The cognition gate evaluates the accumulated streaming memory, while response generation uses a local sliding window containing the most recent codec-token segments. A text query may be inserted at any time, whereas proactive responses are controlled by the cognition gate. This design supports continuous perception, event-triggered commentary, and user-initiated interaction within the same model.

4.2 Data Construction

Our training corpus contains approximately **350M image–caption pairs**, **54M image-instruction samples**, **7.95M unique video–caption samples**, and **3.35M streaming samples**. Caption data establish broad visual-language and temporal grounding, whereas instruction data provide task-oriented reasoning, response formatting, and grounding behavior. The streaming corpus further supervises response timing over causal video prefixes. During video training, selected image-

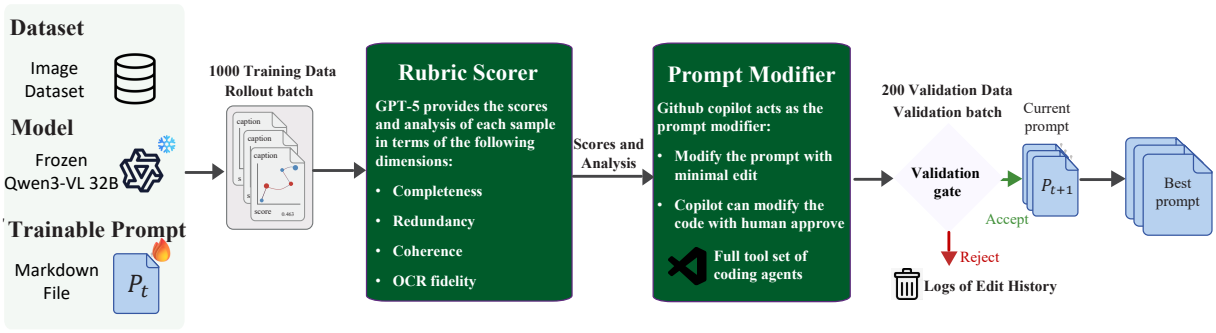

Figure 4 Iterative prompt-optimization pipeline for image recaptioning. A frozen Qwen3-VL-32B generates captions for a rollout batch, which are evaluated by a GPT-5 Rubric Scorer for completeness, redundancy, coherence, and OCR fidelity. GitHub Copilot then proposes minimal prompt or code revisions, which are evaluated on a held-out validation batch and accepted only after human review.

instruction samples are interleaved with video data to preserve static-image understanding and general instruction-following capabilities.

4.2.1 Image Data

Scaled image-caption corpus. We construct approximately 350M image–caption pairs from two complementary sources. The first contains 85M images from the LLaVA-OneVision-1.5 mid-training corpus [8], while the second is obtained by filtering approximately one billion publicly available image–text pairs [93] to retain 265M high-quality images. Images from both sources are recaptioned using the optimized pipeline illustrated in Fig. 4. Compared with short web alt-text, the resulting dense captions provide substantially richer descriptions of foreground and background objects, attributes, counts, spatial relationships, actions, and global scene context. For documents, charts, tables, posters, and screenshots, the captions additionally preserve rendered text, layout structure, and visual–textual relationships, with visible text retained verbatim in its original language whenever possible. Together, these data provide the primary early visual-language supervision for Mage-VL, covering object recognition, attributes, OCR, document understanding, chart interpretation, and spatial relationships before instruction tuning.

Caption-prompt optimization. Driven by the recent rise of the **AI4AI** paradigm—where automated systems and agentic workflows are deployed to iteratively scale and refine data quality [94]. We proactively designed and implemented a closed-loop data optimization pipeline over six months ago during the early development of our framework.

The motivation is that the quality of recaptioned data depends strongly on the system prompt used by the captioning model. We therefore develop an iterative prompt optimization pipeline that combines multi-agent evaluation with human-in-the-loop verification, as illustrated in Fig. 4. The optimized prompt (P_{t+1}) is subsequently used with a frozen Qwen3-VL-32B [5] to generate the large-scale dense-caption corpus.

We initialize the optimization process with a rollout batch of 1,000 training images uniformly sampled from ten representative visual domains, with 100 images from each domain. These domains cover diverse content distributions, including natural scenes, people, objects, documents, charts, rendered text, artworks, screenshots, and other visually structured content. Initial captions are generated using the baseline LLaVA-OneVision-1.5 recaptioning prompt [8].

For each optimization iteration, a rubric-scoring agent powered by GPT-5 evaluates the generated captions along four dimensions:

- **Completeness:** whether the caption covers the major objects, attributes, actions, spatial

relationships, and contextual information visible in the image;

- **Redundancy:** whether the caption contains repeated, verbose, or semantically duplicated descriptions;
- **Coherence:** whether the caption is logically organized, grammatically consistent, and easy to follow;
- **OCR fidelity:** whether visible text is correctly transcribed, preserved in its original language, and associated with the appropriate visual region or document structure.

Based on the scores and diagnostic analysis, a prompt-refinement agent, implemented via GitHub Copilot equipped with a full toolset of coding agents, proposes modifications to the current prompt (P_t). Guided by the principle of making minimal edits, Copilot refines the prompt within a Markdown file. The proposed prompt is then evaluated through a validation gate using a separate validation batch of 200 images. A human reviewer oversees this process to approve or reject the modification. Revisions that fail the validation gate are rejected and archived into the logs of edit history, while approved revisions are accepted as the new baseline prompt (P_{t+1}) for the next iteration. We repeat this evaluate–refine–verify cycle for ten iterations, yielding a final prompt that balances visual coverage, language conciseness, structural coherence, and OCR preservation.

The optimized system prompt is then used for large-scale image captioning. As illustrated in Section 6.1, captions generated with the optimized prompt consistently improve performance across OCR, document, chart, perception, and general visual-reasoning benchmarks. The complete recaptioning system prompt is provided in Appendix A.

Image instruction data. In addition to dense image captions, we collect approximately 54M image-instruction samples. The filtered mixture contains approximately 44.3M samples derived from LLaVA-OneVision-1.5 [8] and FineVision [95], 6.0M samples from OpenBee [96], and 4.6M spatial and GUI instruction samples from LLaVA-OneVision-2 [20]. The mixture covers general visual question answering, OCR, documents, charts, STEM and mathematical reasoning, multi-image understanding, spatial grounding, GUI understanding, and structured multimodal reasoning. The spatial subset emphasizes object size, count, relative direction, distance, depth, and appearance order, while the GUI subset covers interface-element recognition, localization, and action-oriented reasoning. During later video stages, subsets of these samples are interleaved with video-caption and video-instruction data. This mixed training preserves image reasoning and instruction-following capabilities as the temporal context is progressively extended.

4.2.2 Video Data

Our video-caption corpus [20] contains approximately **7.95M unique samples**: 4.2M videos of up to 30 seconds, 2.7M videos spanning 30–60 seconds, 0.7M videos spanning 60–180 seconds, and 350K videos of approximately 10–15 minutes. For the 350K long-video subset, we divide each video into non-overlapping segments and estimate content density using codec-stream bit counts, retaining the segment with the highest bit count as the most visually dynamic interval. This subset is initially processed through standard frame sampling and later revisited using codec-stream tokenization, but is counted only once in the unique-sample total. We additionally incorporate LLaVA-Video-178K [97], TimeLens [98], VideoChat-Flash-Training-Data [43], Molmo2-VideoTrack, and Molmo2-VideoPoint [99] for video question answering, temporal grounding, tracking, point-based localization, and spatio-temporal reasoning.

Duration-stratified video captions. The 4.2M short videos are sampled at 1 fps with at most 30 frames and primarily describe local actions, brief interactions, and object-state changes. The 30–60-

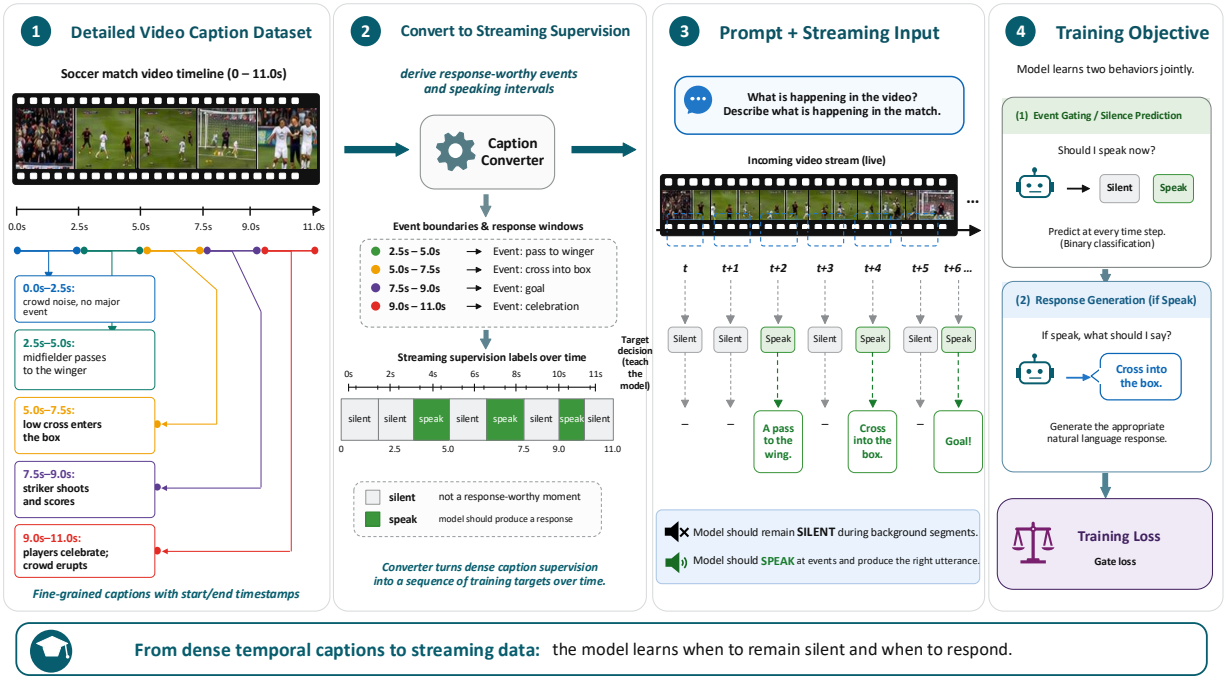

Figure 5 Constructing proactive streaming supervision from timestamped video captions. Timestamped captions are converted into gate and response targets over causal video prefixes. At each caption start timestamp, the target is *speak* and the corresponding caption provides the response; all other candidate timestamps are labeled *silent*. This supervision teaches the model when to remain silent and when to generate an event-conditioned response.

second and 60–180-second subsets use up to 60 and 90 frames, respectively, providing supervision for multi-step activities, scene transitions, event ordering, and longer action sequences. The 10–15-minute subset initially uses up to 384 sampled frames and captures long-range event progression, repeated actions, scene continuity, and cross-segment dependencies. Caption granularity is adapted to video duration. Short clips receive compact, action-dense descriptions, whereas longer videos use segment-level captions with explicit temporal progression and cross-segment references. This supervision teaches the model how visual states evolve over time rather than treating a video as an unordered collection of frames.

Frame-sampled and codec-stream representations. All 7.95M video-caption samples are introduced through conventional frame sampling during the early and intermediate training stages. For codec-native adaptation, the 350K long-video subset is converted into temporally ordered codec windows using the patchifier in Fig. 2. Anchor-frame patches are retained densely, whereas predicted frames contribute only patches selected by codec-derived motion and residual importance. We use configurations of up to 384 and 768 frames. This conversion preserves the original caption supervision while changing the visual observation interface. Codec-stream tokenization allocates tokens according to spatio-temporal variation rather than assigning a dense patch grid to every sampled frame, enabling denser temporal coverage under a comparable visual-token budget.

4.2.3 Streaming Event Data

We construct streaming supervision from our 180-second caption corpus and existing timestamped datasets, including MatchTime [78], LiveCC [77], and StreamingVLM [75]. All sources are converted into a unified real-time *speak*/*silent* format.

Our streaming corpus contains **3.3M training samples**: 2.8M samples converted from our 180-second caption data and 0.5M samples derived from existing timestamped streaming datasets.

Each sample consists of a causal video prefix, a binary *speak*/*silent* target, and (for positive instances) a corresponding language response. All sources are normalized into the same real-time annotation format.

As illustrated in Figure 5, an annotated video is represented as

$$\mathcal{V} = \{(s_j, e_j, c_j)\}_{j=1}^{N},$$

where s_j and e_j denote the start and end timestamps of the j-th segment, and c_j describes the corresponding event, action, or speech. Following the event-gated formulation of StreamMind [100], a caption converter transforms these timestamped annotations into real-time gate and response targets.

For each caption (s_j, e_j, c_j), the start timestamp s_j is assigned a positive *speak* label and paired with c_j as the target response. Candidate timestamps that do not begin a caption receive a negative *silent* label. Applying this conversion to the 180-second caption corpus produces 2,801,360 real-time training samples.

All examples satisfy a *no-future-frame* constraint: the gate decision and target response depend only on observations available up to the current timestamp. Applying the same conversion to the existing timestamped datasets produces the remaining 543,943 samples and increases the diversity of events, response styles, and interaction timings.

4.2.4 Data Curation and Balancing

All data mixtures are filtered for corrupted files, unsafe content, low-quality samples, and near-duplicates. For caption data, we additionally remove captions that are too short to provide meaningful visual grounding, captions whose timestamps are inconsistent with the video duration, repeated or near-duplicate temporal descriptions, and samples flagged by caption–video consistency checks. For streaming supervision, we balance the *speak*/*silent* labels produced by the caption converter by subsampling redundant silent timestamps.

We balance the corpus along two additional axes. Domain balancing upweights regimes that are underrepresented in web data but important for evaluation and deployment, including OCR, documents, charts, STEM and mathematics, spatial reasoning, temporal grounding, and streaming events. Duration balancing controls the proportions of short, medium, long, and ultra-long videos at each training stage. Overall, the early stages inherit the open-data family of LLaVA-OneVision-style training, but place greater emphasis on scaled dense image captions, temporally precise video captions, codec-stream supervision, and calibrated cognition-gate targets.

4.3 Progressive Training Curriculum

Mage-VL is trained through a progressive five-stage supervised curriculum. The five stages successively establish visual-language grounding, instruction following, temporal-context extension, codec-native long-context adaptation, and proactive streaming alignment. Together, they produce a single unified model that supports image understanding, offline video reasoning, and proactive interaction over continuous video streams.

Stage 1: Multimodal alignment via captions. Stage 1 uses approximately 350M dense image captions and 4.2M short-video captions [20]. Dense image captions provide the dominant supervision, forcing the model to associate Mage-ViT tokens with objects, attributes, OCR text, charts, documents, spatial relationships, and global scene context. Fine-grained short-video captions are included from the beginning so that the model encounters motion and state changes before the long-video curriculum. Unlike a projector-only alignment stage, this warm-up directly trains the model as a caption-capable multimodal language model.

Stage 2: Instruction tuning and short temporal grounding. Stage 2 uses approximately 54M image-instruction samples from LLaVA-OneVision-1.5 [8], FineVision [95], OpenBee [96], and LLaVA-OneVision-2 spatial and GUI data [20], together with 3.4M video captions spanning 30–180 seconds [20]. We increase the weight of multimodal instruction data, particularly OCR, document QA, chart reasoning, STEM and mathematics, multi-image reasoning, spatial grounding, and GUI understanding. Short-video captions and instructions remain active to provide local temporal grounding for actions and object-state transitions. Generic language-only or weakly visual SFT is assigned a low weight so that it does not dominate the visual grounding acquired in Stage 1. This stage transforms the model from a visual describer into a general multimodal assistant while retaining caption supervision as regularization.

Stage 3: Temporal-horizon expansion. Stage 3 retains 20M Stage-2 image-instruction samples and introduces medium- and long-video data from LLaVA-Video [97], TimeLens [98], VideoChat-Flash [43], Molmo2 [99], and our temporally structured captions. These data describe event order, scene transitions, repeated actions, and cross-segment dependencies. The goal is not simply to expose the model to more frames, but to teach it how visual states evolve over minutes and how later events depend on earlier context. Image SFT, spatial data, GUI data, and short-video samples remain in the mixture to preserve static perception and instruction-following capabilities. **We also employ AI4AI at this stage. Specifically, we employ AI-based diagnostics to pinpoint underperforming or missing capabilities, which are subsequently addressed by targeted human data supplementation. Meanwhile, these diagnostics dynamically guide the selection of the optimal resolution and frame count.**

Stage 4: Codec-native long-context adaptation. Stage 4 uses 350K long-video captions, 4M spatial samples [20], Molmo2 tracking and pointing data [99], and 40M retained image-instruction samples. After the model has learned to reason over extended temporal contexts, we convert a large fraction of video training into the codec-native representation used at inference time. Captioned videos are represented as rolling token windows containing dense anchor-frame evidence and sparse predicted-frame updates. This stage adapts the language model to the irregular visual sequences produced by Mage-ViT, where token density follows visual change rather than fixed frame slots. Long and ultra-long videos benefit most from this representation because the model can observe more event boundaries under the same visual-token budget. Non-video instruction data remain in the mixture to preserve image, OCR, chart, spatial, and GUI capabilities.

Stage 5: Proactive streaming alignment via cognition-gate fine-tuning. In Stage 5, we introduce a lightweight cognition gate to enable proactive streaming interaction while keeping the base language model completely frozen. For each causal segment, the codec-extraction pipeline packs codec-selected patches into a compact canvas while preserving their original spatio-temporal coordinates. These features are processed by the event-preserving feature extractor (EPFE), whose recurrent state maintains a streaming perception memory $\mathcal{M}_{\text{per}}$ used exclusively by the cognition gate for real-time turn-taking decisions. Crucially, when the gate issues a *speak* decision, we do not employ complex long-range memory for language generation. Instead, the model invokes the frozen base VLM directly using a local sliding window of the most recent N codec-assembled visual segments as input context to execute standard autoregressive response generation. We train the cognition gate on approximately 3.35M streaming samples: 2.8M converted from 180-second captions and 0.54M converted from MatchTime [78], LiveCC [77], and StreamingVLM [75], as described in Section 4.2.3.

Throughout this stage, the visual backbone, EPFE, and language model remain strictly frozen, and we solely fine-tune the cognition gate. At every candidate timestamp, the gate autoregressively evaluates the accumulated streaming memory $\mathcal{M}_{\text{per}}$ and predicts one of two tokens: *silent* or

Table 1 Image and video representation quality of Mage-ViT. Image classification is evaluated using linear probes with 256 visual tokens per image. Video recognition is evaluated using attentive probes with a fixed budget of 4096 visual tokens per video. Chunk-wise evaluation uniformly samples 16 frames with 256 tokens per frame, whereas codec-based evaluation distributes the same budget over 64 frames using sparse patch selection. Baseline video encoders are evaluated using their 16-frame representations. All values denote top-1 accuracy (%).

Image Benchmarks

Method	Backbone	Mode	DTD [101]	CIFAR-10 [102]	SUN397 [103]	Food-101 [104]	ImageNet [105]
SigLIP [27]	ViT-Large/16	Image	85.90	98.24	80.64	95.15	84.46
MetaCLIP2 [106]	ViT-Large/14	Image	85.48	98.76	79.75	93.93	82.74
AIMv2 [107]	ViT-Large/14	Image	86.28	98.95	81.12	94.76	85.26
SigLIP2 [12]	ViT-Large/16	Image	85.90	98.31	**82.36**	**95.85**	**85.92**
DINOv3 [108]	ViT-Large/14	Image	**86.81**	99.17	80.11	94.96	85.38
MoonViT [109]	ViT-SO400M/14	Image	82.77	96.81	77.83	88.56	77.18
OV-Encoder [18]	ViT-Large/14	Image	85.48	99.06	80.60	94.79	84.54
Mage-ViT	ViT-Large/16	Image	85.96	**99.33**	82.01	95.60	85.69

Video Benchmarks

Method	Backbone	Mode	Diving-48 [110]	HMDB-51 [111]	Perception [112]	Charades-Ego [113]	K400 [114]
SigLIP [27]	ViT-Large/16	Chunk	54.70	78.80	51.00	11.70	79.10
MetaCLIP2 [106]	ViT-Large/14	Chunk	42.10	78.20	51.10	11.20	84.00
AIMv2 [107]	ViT-Large/14	Chunk	55.70	82.60	56.40	12.40	82.20
SigLIP2 [12]	ViT-Large/16	Chunk	62.30	84.61	58.47	13.75	83.93
DINOv3 [108]	ViT-Large/14	Chunk	61.30	79.70	**60.80**	**14.00**	83.90
MoonViT [109]	ViT-SO400M/14	Chunk	43.95	77.17	56.95	10.56	79.73
OV-Encoder [18]	ViT-Large/14	Chunk	60.86	83.68	60.04	12.61	84.81
		Codec	**65.32**	84.80	60.16	12.81	**84.87**
Mage-ViT	ViT-Large/16	Chunk	60.45	85.13	58.91	13.49	84.83
		Codec	64.14	**85.17**	59.17	13.31	84.66

speak. Caption start timestamps provide speak targets, while all other timestamps serve as silent targets. To address the inherent sparsity of speak events, we optimize a class-weighted token cross-entropy loss:

$$\mathcal{L}_{\text{gate}} = -\sum_{t} w_{g_t} \log p(g_t \mid g_{<t}, \mathcal{M}_{\text{per}}), \qquad g_t \in \{\text{silent}, \text{speak}\}.$$

When g_t = speak, the recent N-segment codec visual canvas is passed directly to the frozen language model for autoregressive generation. Restricting optimization to the cognition gate adds proactive response timing while preserving the image and video understanding capabilities acquired during the preceding stages.

5 Experiments

5.1 Experimental Setup

Models under evaluation. We evaluate Mage-VL-4B, which combines the from-scratch Mage-ViT visual tokenizer (Section 3) with a Qwen3-4B-Instruct-2507 language backbone [92]. At inference time, videos are represented as codec-token streams on a shared 16×16 patch grid: anchor frames are encoded densely, whereas predicted frames contribute only patches selected by the codec-derived importance signal. We report three codec-canvas operating points with different accuracy–efficiency trade-offs. The default *tc32* setting uses the largest visual budget and generally delivers the strongest performance; *tc16* provides a balanced configuration with lower evaluation cost and competitive accuracy; and the lightweight *tc8* setting prioritizes latency while retaining much of the higher-budget performance. Their nominal canvas budgets correspond to the visual workloads of 32, 16, and 8 uniformly sampled frames, respectively, enabling matched-budget comparisons with frame-sampling baselines. Still images use single-image spatial patchification.

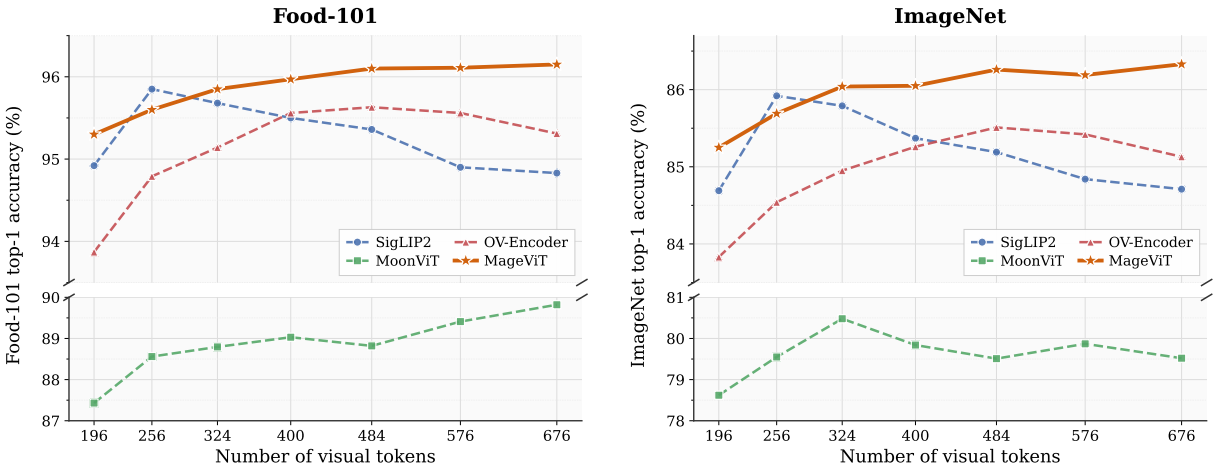

Figure 6 Effect of the visual-token budget on image representation quality. We evaluate frozen visual features on Food-101 and ImageNet while increasing the number of input visual tokens. Mage-ViT, which is trained with variable-resolution inputs, consistently benefits from larger token budgets and achieves its best performance at the highest evaluated resolution. In contrast, the fixed-resolution baselines generally saturate or degrade after reaching their optimal token budget.

Baselines. Our primary comparison is Qwen3-VL-4B-Instruct [5], which uses a language model of the same parameter scale but a conventional dense visual front-end. We additionally compare with Phi-4-Multimodal-Instruct (Phi-4-MM, 5.6B) [21] and Phi-4 reasoning-vision (Phi-4-R-V, 15B) [22]. For the visual-representation study in Section 5.2, we compare Mage-ViT with public contrastive and self-supervised encoders, including SigLIP [27], SigLIP2 [12], MetaCLIP2 [106], AIMv2 [107], DINOv3 [108], MoonViT [109], and OV-Encoder [18].

Benchmarks. We organize the evaluation into four groups. First, representation quality (Table 1) directly evaluates Mage-ViT using linear probes for image classification and attentive probes for video recognition. Second, image understanding (Table 3) covers document and chart understanding, OCR, general VQA, and 2D/3D spatial intelligence. Third, video understanding (Tables 4 and 5) covers video QA, temporal grounding, video spatial reasoning, and referring-video tracking. Finally, proactive streaming is evaluated in Section 5.4 in terms of both response timing and response quality.

Protocol. All supported downstream benchmarks are evaluated using `lmms-eval` [115] with the official prompt templates and scoring rules. We apply no model-specific prompt tuning, semantic rewriting, or manual output correction. For matched-budget video comparisons, the visual-input budget is controlled as described in Section 5.5. Temporal-grounding and tracking outputs are parsed using the official evaluation procedures. Wall-clock measurements in Table 5 are collected on a single node with eight B200 GPUs.

5.2 Evaluation of Mage-ViT

We evaluate Mage-ViT from three perspectives: representation quality under limited pre-training data, the effect of native-resolution training across visual-token budgets, and robustness to different codec families. Image representations are evaluated using linear probes, while video representations are evaluated using attentive probes. Unless otherwise specified, all visual encoders remain frozen.

Table 2 Cross-codec robustness of codec-guided patch selection. We compare the HEVC selector used during training with the neural codec DCVC-RT at inference time, without codec-specific retraining. We report the task score and the average token Canvas per video. DCVC-RT is constrained to use no more tokens than HEVC.

	Traditional: HEVC [81]		Neural: DCVC-RT [82]	
Benchmark	Score	Canvas	Score	Canvas
Video QA				
MV-Bench [116]	65.1	34.0	66.5	30.5
NextQA [117]	83.1	32.7	82.6	32.1
TempCompass [118]	62.3	34.0	61.7	31.1
VideoMME [119]	64.0	33.2	64.3	30.2
LongVideoBench [120]	61.3	32.6	60.0	29.3
LVBench [121]	41.8	32.2	43.5	29.6
MLVU-dev [122]	68.7	33.6	66.1	29.9
Temporal grounding				
Charades [123]	31.4	33.7	31.6	32.0
Timelens-Charades [124]	50.7	33.5	50.5	31.9
Timelens-ActivityNet [124]	45.4	33.7	44.2	31.4
Timelens-QVHighlight [124]	57.4	34.3	57.5	30.7
Spatial reasoning				
VSI-Bench [125]	64.3	32.8	63.7	31.3
Avg	**58.0**	**33.4 (100%)**	**57.7**	**30.8 (92%)**

5.2.1 Learning Strong Visual Representations from Limited Data

As summarized in Table 1, we comprehensively evaluate Mage-ViT on both standard image recognition benchmarks via linear probing and video action/event recognition via attentive probing. Across both modalities, Mage-ViT demonstrates strong representation capacities without requiring massive web-scale pre-training datasets.

Image representation performance. On standard 2D image benchmarks, Mage-ViT delivers performance that consistently rivals or exceeds top-tier visual encoders trained on billions of images. Specifically, Mage-ViT achieves the highest top-1 accuracy on CIFAR-10 (**99.33%**) and exhibits highly competitive performance on coarse-grained and fine-grained classification tasks, such as SUN397 (**82.01%**), Food-101 (**95.60%**), and ImageNet-1K (**85.69%**). Notably, its performance closely matches SigLIP2 (**85.92%** on ImageNet) and surpasses established strong backbones including MetaCLIP2, AIMv2, and OV-Encoder. Importantly, while OV-Encoder relies on a smaller 14×14 patch size, Mage-ViT adopts a larger 16×16 patch grid, achieving comparable or superior representation quality with fewer visual tokens. This confirms that our pre-training scheme effectively constructs a highly transferable and semantically rich feature space for static visual concepts using significantly fewer visual tokens and pre-training samples.

Video representation performance and temporal adaptation. When extended to temporal video understanding, Mage-ViT demonstrates superior temporal modeling and efficiency, particularly in motion-heavy and long-horizon scenarios:

- **Chunk-wise Evaluation:** Under the standard 16-frame uniform sampling protocol, Mage-ViT achieves a top-1 accuracy of **85.13%** on HMDB-51 and **84.83%** on Kinetics-400 (K400), outperforming strong baselines such as SigLIP (**78.80%** and **79.10%**) and DINOv3 (**79.70%** and **83.90%**). On fine-grained action perception benchmarks like Diving-48 (**60.45%**), Mage-ViT

significantly outperforms traditional image-centric models, highlighting its robust spatial-temporal feature extraction.

- **Codec-based Sparse Sampling:** When leveraging sparse patch selection across an expanded temporal context of 64 frames under the exact same visual token budget (4096 tokens), Mage-ViT exhibits substantial performance boosts on motion-critical tasks. Specifically, the accuracy on Diving-48 jumps from **60.45%** to **64.14%** (+3.69%), and HMDB-51 reaches a peak accuracy of **85.17%**. This confirms that the spatio-temporal representations learned by Mage-ViT naturally pair with codec-driven sparse tokenization, enabling the model to capture fine-grained temporal dynamics over longer video horizons without increasing computational or memory overhead.

> ***Finding 1:*** **Web-Scale Pre-training is Not Essential for VLM Visual Encoders.** Despite being trained from scratch on a modest corpus, Mage-ViT achieves visual representation quality competitive with flagship encoders like SigLIP2, proving that architectural and objective design can effectively substitute for web-scale pre-training data.

5.2.2 Native-Resolution Scaling across Visual-Token Budgets

To further evaluate how visual representations adapt to flexible input resolutions, we analyze performance across an increasing visual token budget (from 196 to 676 tokens per image) on Food-101 and ImageNet-1K, as shown in Fig. 6.

Monotonic resolution scaling vs. fixed-resolution degradation. Visual encoders exhibit fundamentally distinct scaling behaviors depending on their pre-training resolution strategies:

- **Fixed-Resolution Baselines (SigLIP2 & OV-Encoder):** Because these models are pre-trained on fixed spatial grids, evaluating them under larger token budgets introduces severe positional distribution shifts. Consequently, SigLIP2 reaches its peak performance at 256 tokens on ImageNet ($\sim$ 85.9%) and continuously degrades as the token budget expands to 676 ($\sim$ 84.7%). Similarly, OV-Encoder experiences a clear performance drop beyond 484 tokens.
- **Variable-Resolution Encoders (MoonViT & Mage-ViT):** Encoders pre-trained with variable-resolution objectives naturally adapt to flexible spatial formats. While MoonViT avoids high-resolution collapse, its overall representation capacity remains low across all budget settings. In contrast, Mage-ViT achieves the best of both worlds: its representation quality monotonically improves as the visual token budget increases—reaching peak accuracy at 676 tokens ($>$ 96.1% on Food-101 and $>$ 86.3% on ImageNet)—while consistently outperforming all baselines at every evaluated resolution.

This superior extrapolation capacity confirms that Mage-ViT seamlessly accommodates dynamic visual sequence lengths, making it an ideal visual backbone for Large Multimodal Models (LMMs).

> ***Finding 2:*** **Variable-Resolution Pre-training Enables Continuous Monotonic Scaling with Resolution and Token Budgets.** While fixed-resolution vision encoders suffer from sharp performance degradation when evaluated at higher resolution budgets, Mage-ViT leverages variable-resolution pre-training to achieve continuous, monotonic quality gains without saturation.

Table 3 Image understanding and spatial intelligence across STEM/math, document understanding, general VQA, perception/alignment, and 2D/3D scene geometry, embodied/viewpoint, cross-view, and relational spatial benchmarks. Dark blue and light blue background colors indicate the best and second-best results per row, respectively.

Benchmark	Mage-VL-4B	Qwen3-VL-4B [5]	Phi-4-MM-5.6B [21]	Phi-4-R-V-15B [22]
Document understanding				
DocVQA-val [126]	**95.14**	94.69	92.79	76.20
InfoVQA-val [127]	**80.33**	79.50	71.84	55.41
AI2D w/ Mask [128]	**83.16**	81.54	81.83	82.87
AI2D w/o Mask [128]	91.87	92.20	91.35	**93.81**
ChartQA [129]	**84.88**	83.96	83.76	83.40
OCRBench [130]	**81.80**	81.60	81.70	73.90
CC-OCR Doc [131]	32.25	**39.69**	4.99	17.65
MultiDocVQA-val [132]	**87.46**	87.21	46.84	58.35
DUDE [133]	46.44	**50.98**	3.78	34.34
WebSRC-val [134]	92.80	**95.40**	71.30	76.60
ChartQAPro [135]	**32.57**	26.79	0.13	25.38
TextVQA-val [136]	77.28	**80.55**	39.93	76.06
CharXiv-Description [137]	75.85	74.40	18.65	**75.98**
CharXiv-Reason [137]	35.20	26.10	0.10	**36.10**
General VQA				
MMBench-EN-dev [138]	84.02	83.25	65.81	**84.19**
MMBench-CN-dev [138]	**82.04**	80.58	75.17	79.47
RealWorldQA [139]	70.46	**70.85**	60.65	70.72
MMStar [140]	**67.32**	62.04	61.24	59.63
MME-Perception [141]	**1709.54**	1703.50	1409.66	1590.21
SeedBench (All) [142]	**76.78**	75.65	68.28	73.70
SeedBench-Image [142]	**79.32**	78.87	74.48	77.97
SeedBench2-Plus [143]	69.21	**69.65**	68.38	69.48
MMT-val [144]	64.47	**66.52**	59.19	64.89
CV-Bench [10]	**87.79**	85.37	57.09	81.31
MME-RealWorld [145]	**66.52**	63.20	32.45	57.80
MME-RealWorld-CN [145]	61.33	**63.09**	21.19	46.07
Spatial intelligence				
CV-Bench-2D [146]	**82.13**	81.00	56.12	80.11
CV-Bench-3D [146]	**94.75**	92.30	56.92	82.50
BLINK [147]	**65.11**	65.10	35.24	57.80
EmbSpatial [148]	**82.67**	77.50	41.51	72.67
CRPE-Relation [149]	76.12	**77.70**	34.60	74.46
CrossPoint [150]	**80.00**	26.90	12.20	47.73
ERQA [151]	36.00	**42.30**	30.75	40.25
MMSI-Bench [152]	28.20	**31.00**	28.80	25.70
SAT [153]	67.33	**69.30**	55.33	66.67

5.2.3 Robust Token Selection across Codec Families

As introduced in Section 3.1, Mage-ViT uses a codec-derived importance map to select informative patches from predicted frames. During training, this map is computed from HEVC [81] using motion-vector magnitude and residual energy. Since these signals are specific to the coding process of a traditional codec, an important question is whether the resulting token-selection strategy generalizes to fundamentally different codec families.

We evaluate this by replacing the HEVC-based selector at inference time with DCVC-RT [82], a neural video codec whose learned probability model estimates local coding cost through negative log-likelihood. No codec-specific retraining or adaptation is applied. As shown in Table 2, the two selectors achieve nearly identical average performance across video QA, temporal grounding, and spatial reasoning benchmarks, while the neural-codec selector uses a smaller average token canvas. Performance also remains comparable across most individual tasks, with no consistent degradation after switching codec families.

These results indicate that codec-guided token selection does not depend on the particular motion-estimation, residual-coding, or probability model used by a codec. Although HEVC and DCVC-RT implement compression through different mechanisms, both assign greater coding cost

Table 4 Video understanding and temporal grounding. Mage-VL and Qwen3-VL-4B share the 4B Qwen3 LLM backbone and differ only in the visual front-end (Qwen3-VL: dense ViT; Mage-VL: Mage-ViT); Phi-4-MM and Phi-4-R-V are reported for reference. Dark blue and light blue background colors indicate the best and second-best results per row, respectively.

Benchmark	Mage-VL-4B	Qwen3-VL-4B [5]	Phi-4-MM-5.6B [21]	Phi-4-R-V-15B [22]
Video QA				
MV-Bench [116]	65.1	**66.7**	44.9	49.2
NextQA [117]	**83.1**	79.8	54.1	69.0
VideoMME [119]	**64.0**	59.7	44.7	55.3
LongVideoBench [120]	**61.3**	57.7	41.14	51.2
LVBench [121]	**41.8**	39.2	25.31	34.4
MLVU-dev [122]	**68.7**	61.5	44.18	51.8
VideoMME (w/ sub.) [119]	66.3	**70.2**	45.4	58.3
VideoMME-V2 [154]	24.3	**24.4**	19.2	23.9
VideoEval-Pro [155]	**45.2**	20.7	14.35	16.8
JumpScore [20]	**45.60**	3.82	3.61	1.53
MMOU-Test-Mini [156]	39.3	38.1	32.22	**51.9**
Temporal grounding				
Timelens-Charades [124]	**50.7**	43.1	4.09	20.6
Timelens-ActivityNet [124]	**45.4**	28.4	2.03	23.0
Timelens-QVHighlight [124]	**57.4**	34.9	2.47	11.6
Spatial reasoning				
VSI-Bench [125]	**64.3**	53.3	24.09	25.5
Tracking (J&F)				
Ref-DAVIS17 [157]	**25.83**	7.48	3.14	2.15
MeViS-ValidU [158]	**22.55**	3.16	10.28	1.53
ReasonVOS [159]	**17.76**	9.66	9.50	9.77
Ref-YT-VOS [160]	**25.57**	5.28	8.64	3.85

to regions that are difficult to predict from temporal context. Motion-vector magnitude, residual energy, and neural-codec negative log-likelihood therefore provide different estimates of the same underlying signal: local temporal predictability.

Importantly, Mage-ViT consumes only the resulting patch-importance map, rather than codec-specific syntax or latent representations. The visual tokenizer is therefore decoupled from the mechanism used to estimate coding difficulty. This establishes temporal predictability as a transferable criterion for allocating visual tokens and allows codec-guided tokenization to generalize across traditional and neural compression systems.

5.3 Evaluation of Mage-VL

5.3.1 Image Understanding

Table 3 reports model performance across three core capability dimensions: document understanding, which aligns with Microsoft's primary domain applications; general VQA, which evaluates general multimodal proficiency; and spatial intelligence, which represents a core strategic focus of Mage-VL. Against the matched-LLM Qwen3-VL-4B control, Mage-VL is competitive on aggregate and clearly ahead on the categories that our caption-heavy curriculum targets. On document, OCR, and chart understanding, Mage-VL leads on the majority of benchmarks—including DocVQA (95.14), InfoVQA (80.33), ChartQA (84.88), OCRBench (81.80), and AI2D-with-mask (83.16)—consistent with the dense-recaptioning and verbatim-OCR emphasis of our data pipeline. On general VQA, the two models trade wins within small margins, with Mage-VL achieving top results on MMStar (67.32), MME-Perception (1709.54), and CV-Bench (87.79), while comprehensively surpassing the similarly-sized Phi-4-MM across the board. The clearest and most consistent gains appear in the spatial-intelligence block: Mage-VL improves over Qwen3-VL on

Table 5 Video performance and evaluation efficiency across token/codec budgets. Performance uses each benchmark's primary metric (↑); evaluation efficiency is reported as wall-clock Time in seconds (↓) on a single 8×B200 node. Qwen3-VL and Phi-4-R-V use 32 frames, and their time excludes estimated video-loading time ($1.01N/8$ and $1.01N$ for N samples, respectively), whereas Mage-VL reports full measured wall-clock time. Dark blue and light blue backgrounds indicate the best and second-best performance per row.

	Mage-VL-4B (tc8)		Mage-VL-4B (tc16)		Mage-VL-4B (tc32)		Qwen3-VL-4B [5]		Phi-4-R-V-15B [22]	
Benchmark	Perf. (↑)	Time (s) (↓)	Perf. (↑)	Time (s) (↓)	Perf. (↑)	Time (s) (↓)	Perf. (↑)	Time (s) (↓)	Perf. (↑)	Time (s) (↓)
Video QA										
MV-Bench [116]	65.1	893	65.5	972	65.1	1054	**66.7**	1268	49.2	2923
NextQA [117]	80.8	415	82.4	502	**83.1**	642	79.8	1460	69.0	29024
TempCompass [118]	61.5	389	62.4	510	62.3	729	**72.7**	433	34.8	3518
VideoMME [119]	57.9	270	61.8	439	**64.0**	534	59.7	463	55.3	3187
LongVideoBench [120]	56.2	279	**63.1**	1308	61.3	345	57.7	365	51.2	2617
LVBench [121]	38.9	213	38.9	266	**41.8**	333	39.2	554	34.4	2692
MLVU-dev [122]	63.2	296	65.6	326	**68.7**	361	61.5	786	51.8	6326
Temporal grounding										
Charades [123]	34.1	534	33.2	417	31.4	729	**45.9**	707	0.3	1844
Timelens-Charades [124]	46.4	483	50.4	538	**50.7**	623	43.1	643	20.6	3607
Timelens-ActivityNet [124]	32.9	554	39.7	622	**45.4**	785	28.4	766	23.0	4611
Timelens-QVHighlight [124]	42.6	357	52.1	358	**57.4**	421	34.9	402	11.6	1643
Spatial reasoning										
VSI-Bench [125]	55.7	329	61.1	368	**64.3**	537	53.3	255	25.5	4360

CV-Bench-3D (94.75 vs. 92.30), EmbSpatial (82.67 vs. 77.50), and dramatically on cross-view point correspondence (CrossPoint 80.00 vs. 26.90), where codec-aligned patchification preserves the dense, geometrically-consistent structure that viewpoint matching requires. Despite using roughly a quarter of its parameters, Mage-VL also surpasses Phi-4-R-V (15B) on the large majority of these spatial benchmarks.

5.3.2 Video Understanding

Table 4 reports video understanding under the same matched-LLM comparison, so that Mage-VL and Qwen3-VL-4B differ only in the visual front-end. Mage-VL improves on most video-QA benchmarks, including VideoMME (+4.3), MLVU (+7.2), LongVideoBench (+3.5), LVBench (+2.6), NextQA (+3.3), and especially the long and realistic VideoEval-Pro (+24.5). Qwen3-VL remains stronger on MV-Bench, TempCompass, and MMVU, which lean more on dense per-frame appearance than on temporal localization. The gains concentrate in *localization-heavy* regimes: on temporal grounding Mage-VL improves by +7.6, +17.1, and +22.5 on the Timelens Charades/ActivityNet/QVHighlight splits; on video spatial reasoning by +11.0 on VSI-Bench; and on referring-video tracking by roughly +18 to +20 J&F points on Ref-DAVIS17, MeViS, and Ref-YouTube-VOS (and +8.1 on ReasonVOS).

Notably, these competitive long VideoQA capabilities are achieved *without any* long VideoQA SFT; instead, Mage-VL relies solely on short video SFT alongside detailed video captions. The dense temporal-visual alignment established by video captions enables the base LLM to handle VideoQA tasks zero-shot, bypassing the need for heavy VideoQA instruction tuning. Table 5 further shows that the lightweight tc8 setting preserves most of these gains at a fraction of the visual-token cost, confirming that the advantage does not hinge on a large frame budget.

> ***Finding 3:*** **Dense Video Captions Bypass Long VideoQA SFT.** Without explicit long VideoQA instruction tuning, fine-tuning solely on video detailed captions (alongside short video SFT) is sufficient for strong VideoQA capabilities. High-quality captioning establishes solid temporal-visual alignment, allowing the base LLM to generalize to long VideoQA zero-shot.

Beyond temporal grounding and video-level reasoning, we observe an intriguing positive transfer

Table 6 Response timing evaluation on SoccerNet-Caption [161]. Metrics include TriggerAcc, TimVal, per-position F1, ROC-AUC, and PR-AUC. Best results per column are in **bold**.

Method	TriggerAcc	TimVal	F1	ROC-AUC	PR-AUC
StreamMind [100]	52.18	47.36	–	–	–
JoyAI-VL-Interaction-9B [73]	**97.98**	19.25	3.55	56.26	1.68
Mage-VL-4B	79.21	**55.54**	**16.35**	**83.14**	**9.30**

from dynamic video training to static spatial intelligence. While traditional paradigms often treat spatial reasoning and temporal video understanding as isolated capabilities, exposure to continuous video streams—rich in egocentric perspective shifts, camera trajectories, and object-environment interactions—provides a strong inductive bias for fine-grained 2D/3D geometric perception. Although rigorous single-variable ablations remain challenging due to the tightly coupled joint pre-training recipe, our empirical results across spatial benchmarks (e.g., CV-Bench and EmbSpatial) strongly support this cross-task synergy.

> ***Finding 4:*** **Dynamic Video Training Synergizes with Spatial Intelligence.** Incorporating dynamic video sequences with rich viewpoint transitions and spatial interactions enhances 2D/3D spatial reasoning capabilities, allowing Mage-VL to achieve strong performance on static spatial benchmarks alongside temporal video tasks.

5.4 Streaming Video Understanding

To systematically evaluate Mage-VL in continuous video streams, we investigate both its *response timing* (when to speak) and *response quality* (what to say), accompanied by a qualitative case study on real-world sports broadcasts.

Response timing. We evaluate our codec-native streaming model on the SoccerNet-Caption validation split following the StreamMind evaluation protocol [100]. The decision gate is evaluated at every valid streaming position via an argmax over the *silent*/*speak* logits using exact canvas-position matching with zero temporal tolerance. We report TriggerAcc (fraction of correctly predicted gate positions) and TimVal (which balances speak and silence correctness), both macro-averaged across 98 match halves. Additionally, raw per-position F1, ROC-AUC, and PR-AUC are reported without temporal tolerance.

For reference, baseline numbers for StreamMind are directly taken from its original publication [100]. Notably, StreamMind was explicitly trained in-distribution on SoccerNet-Caption, whereas Mage-VL achieves superior event-triggering performance under strict, zero-tolerance canvas-position matching without domain-specific over-tuning. Note that JoyAI is evaluated at 1 Hz with a relaxed ±1-second window.

As presented in Table 6, JoyAI-VL-Interaction exhibits a severe bias toward predicting silence. Due to the extreme class imbalance in SoccerNet-Caption (where silent frames dominate), JoyAI achieves an artificially inflated TriggerAcc (97.98%) but suffers a severe degradation on precision-sensitive and balanced metrics (e.g., 19.25% TimVal and 3.55% F1). In contrast, Mage-VL significantly outperforms StreamMind—despite StreamMind being trained *in-distribution* on SoccerNet-Caption—achieving 79.21% TriggerAcc and a state-of-the-art 55.54% TimVal. This superior performance across ROC-AUC (83.14%) and PR-AUC (9.30%) demonstrates that Mage-VL achieves well-calibrated, timely response triggering without collapsing into trivial silent predictions.

Table 7 Results on OVO-Bench. Baseline results and table structure are adapted from SimpleStream [162]. OVO-Bench reports per-task accuracy under Real-Time Visual Perception and Backward Tracing. †: Qwen2.5-VL-7B + HERMES (4K tokens). OCR, ACR, ATR, STU, FPD, and OJR denote Optical Character Recognition, Action Recognition, Attribute Recognition, Spatial Understanding, Future Prediction, and Object Recognition; EPM, ASI, and HLD denote Episodic Memory, Action Sequence Identification, and Hallucination Detection. "Avg." is the mean of the Real-Time and Backward category averages. Dark blue and light blue backgrounds highlight the best and second-best results across all models, respectively.

Model	#Frames	*OVO-Bench* [80]											
		Real-Time Visual Perception							*Backward Tracing*				Avg.
		OCR	ACR	ATR	STU	FPD	OJR	Avg.	EPM	ASI	HLD	Avg.	
Human	–	94.0	92.6	94.8	92.7	91.1	94.0	93.2	92.6	93.0	91.4	92.3	92.77
Offline Video LLMs													
Qwen2.5-VL-7B [6]	1 fps	67.8	55.1	67.2	42.1	66.3	60.9	59.9	51.5	58.8	23.7	44.7	52.28
LLaVA-Video-7B [163]	64	69.1	58.7	68.8	49.4	74.3	59.8	63.5	56.2	57.4	7.5	40.4	51.95
Qwen3-VL-4B [5]	64	77.9	69.7	77.6	60.1	76.2	**75.5**	72.8	**63.0**	66.2	30.1	**53.1**	63.00
Online / Streaming Video LLMs													
VideoLLM-online-8B [69]	2 fps	8.1	23.9	12.1	14.0	45.5	21.2	20.8	22.2	18.8	12.2	17.7	19.26
Flash-VStream-7B [74]	1 fps	24.2	29.4	28.5	33.7	25.7	28.8	28.4	39.1	37.2	5.9	27.4	27.90
Dispider-7B [71]	1 fps	57.7	49.5	62.1	44.9	61.4	51.6	54.6	48.5	55.4	4.3	36.1	45.35
TimeChat-Online-7B [164]	1 fps	75.2	46.8	70.7	47.8	69.3	61.4	61.9	55.9	59.5	9.7	41.7	51.80
StreamForest-7B [165]	1 fps	68.5	53.2	71.6	47.8	65.4	60.9	61.2	58.9	64.9	32.3	52.0	56.60
Streamo-7B [166]	1 fps	79.2	57.8	75.0	49.4	64.4	70.1	66.0	54.6	52.0	31.7	46.1	56.05
HERMES-7B† [167]	1 fps	85.2	64.2	71.6	53.4	74.3	65.2	69.0	48.5	62.2	**37.6**	49.4	59.20
JoyAI-VL-Interaction-9B [73]	1 fps	72.5	61.5	75.0	59.0	**78.2**	64.1	68.4	62.0	**70.9**	12.9	48.6	58.50
Mage-VL-4B	1 fps	**94.63**	**83.49**	**79.31**	**74.16**	70.30	77.17	**79.84**	50.17	61.49	32.80	48.15	**64.00**

Response quality. Beyond timing accuracy, we evaluate the streaming perception quality of Mage-VL on OVO-Bench [80] under the recent-window protocol established by SimpleStream [162], where queries are answered using the four most recent frames sampled at 1 fps. As detailed in Table 7, Mage-VL achieves 79.84% average accuracy on Real-Time Visual Perception and 48.15% on Backward Tracing, establishing a new state-of-the-art overall OVO-Bench score of 64.00% among streaming architectures. Crucially, this competitive real-time perception capability is attained without requiring streaming-specific fine-tuning or heavy external memory modules.

Qualitative illustration. Fig. 7 presents a qualitative case study on a live 2026 World Cup broadcast between England and Argentina outside the SoccerNet evaluation split. As visualized, our codec front-end dynamically retains spatial patches (16×16) associated with salient motion and novel visual details while suppressing temporally predictable background regions. Throughout routine gameplay (02.16s–24.50s), the event gate stays silent. At 29.36s, upon detecting an active offensive play—where an England player drives toward the goal under close pursuit by an Argentina defender—the event gate triggers the language decoder to proactively generate a real-time commentary (*"An England player drives toward goal, closely pursued by an Argentina defender as the attack develops."*). This demonstrates the seamless synergy between codec-native sparse perception and event-driven response generation in demanding live streaming scenarios.

5.5 Efficiency Analysis

To understand how codec-native tokenization alters the trade-off between visual budget and downstream accuracy, we analyze both the scaling trends under matched visual input budgets and the wall-clock evaluation efficiency.

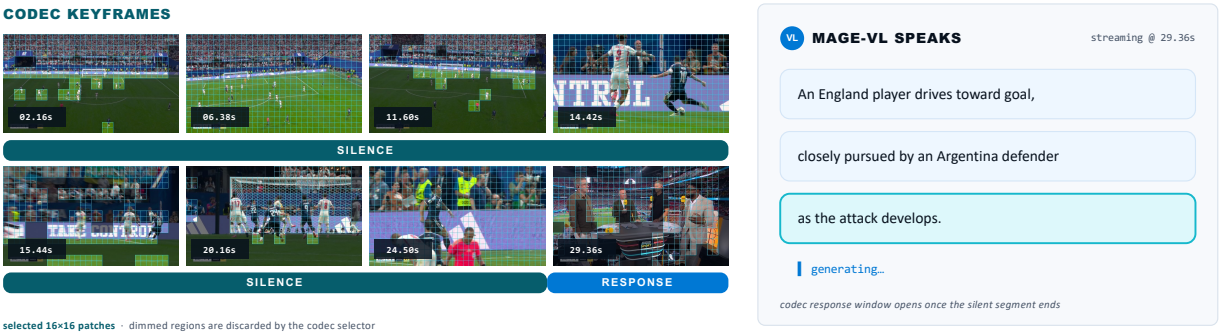

Figure 7 **Codec-native proactive streaming on a 2026 World Cup broadcast (England vs. Argentina).** Eight frames from a causal match window visualize the 16×16 patches retained by the codec selector; dimmed regions are discarded as temporally predictable. The event gate remains silent through routine updates and triggers the language decoder at 29.36s when a response-worthy offensive play is observed. The speech bubble shows the real-time generated response.

Matched-budget setup. We evaluate *codec-input efficiency* by comparing downstream performance under an equivalent nominal visual input capacity to the vision encoder. For the uniform baseline, *frame-N* denotes Qwen3-VL-4B evaluated on $N \in \{4, 8, 16, 32, 64\}$ uniformly sampled RGB frames. For Mage-VL, *tc-N* denotes a target budget of N codec canvases, constructed via adaptive temporal grouping and salient patch packing over an 8× denser source-frame sequence ($8N$ raw frames). Following LLaVA-OneVision-2 [20], both paradigms share identical vision architectures (16×16 patch size, max 150K pixels per input canvas). Consequently, frame-N and tc-N subject the vision backbone to the same nominal token workload, but tc-N adaptively concentrates representation capacity onto motion-salient regions across a significantly wider temporal horizon.

Accuracy-budget scaling. As illustrated in Fig. 8, codec-native tokenization establishes a consistently superior accuracy-efficiency frontier over uniform sampling across long-video and fine-grained temporal tasks. Crucially, the performance advantage (*green shaded area*) is most pronounced in low-to-medium budget regimes ($N \leq 16$). On dense temporal localization benchmarks such as TimeLens-ActivityNet, TimeLens-Charades, and TimeLens-QVHighlights, sparse uniform sampling frequently misses crucial short-duration events, whereas tc-N retains vital temporal evidence even at $N = 4$. On long-video QA benchmarks (e.g., VideoMME, LVBench, and MLVU-Dev), tc-N steadily outscales uniform sampling across all budgets ($4 \leq N \leq 64$). The only notable boundary case occurs on MVBench, where both representations converge at high frame budgets, as its short clip structures offer fewer temporally redundant regions for codec compression to exploit. Overall, codec tokenization acts as a dynamic capacity reallocator—pruning temporal redundancies to maximize information density per visual token.

Pareto efficiency in wall-clock time. Table 5 further highlights the practical wall-clock inference advantages of Mage-VL. By compressing redundant video content prior to vision encoding, Mage-VL achieves superior accuracy while drastically reducing latency. For instance, on VideoMME, Mage-VL under *tc8* achieves 57.9% accuracy in just 270 s wall-clock time—outperforming Qwen3-VL (59.7% at 463 s) in compute time while matching its accuracy at *tc16* (61.8% at 439 s). On NExT-QA [168], Mage-VL (*tc8*) cuts the evaluation wall-clock time from 1460 s down to 415 s (3.5× speedup) while achieving higher accuracy (80.8% vs. 79.8%). These results confirm that codec-native visual representation delivers genuine end-to-end inference speedups alongside strong empirical accuracy gains.

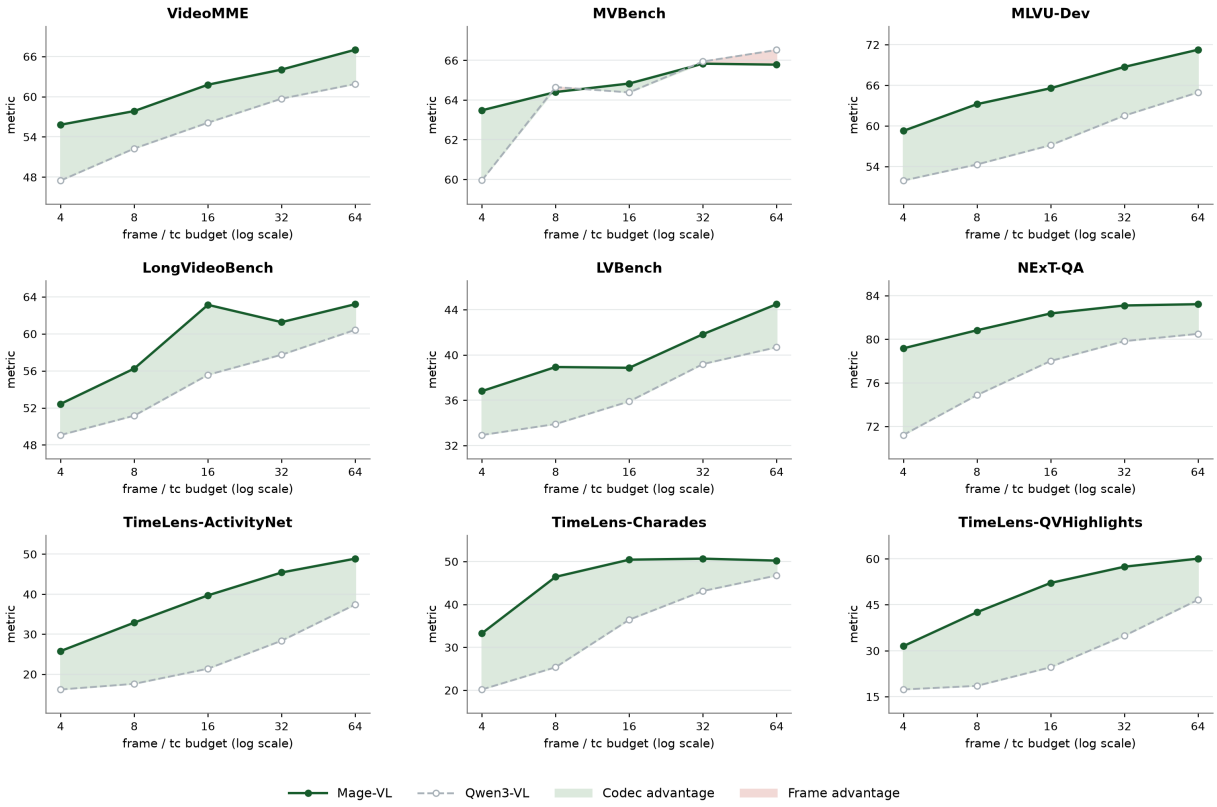

Figure 8 **Codec inputs versus uniform frame sampling across input budgets.** Frame-N uses N uniformly sampled RGB frames, whereas tc-N targets N codec canvases constructed from $8N$ source frames. Shaded regions highlight the performance gain of codec-native tokenization.

Table 8 Downstream performance comparison before and after image captioning pipeline optimization.

Configuration	AI2D [128]	ChartQA [129]	InfoVQA [127]	MMStar [140]	OCRBench [130]	MMBench-CN [138]	DocVQA [126]	CV-Bench [10]	Realworld QA [139]
Before Optimization	82.74	84.80	74.26	62.91	76.90	78.78	93.66	78.51	69.41
After Optimization	**82.97**	**86.32**	**79.88**	**63.43**	**80.70**	**80.15**	**94.74**	**79.68**	**73.20**
Gain (Δ)	+0.23	+1.52	+5.62	+0.52	+3.80	+1.37	+1.08	+1.17	+3.79

Finding 5: **Codec-Native Inputs Significantly Boost Video Representation Efficiency.** By dynamically packing motion-salient patches across extended temporal horizons, codec-native tokenization establishes a superior accuracy-efficiency frontier over uniform frame sampling, achieving substantially higher performance under matched token budgets and up to 3.5× wall-clock inference speedups.

6 Discussions

6.1 AI4AI Data Pipeline

To evaluate the effectiveness of the prompt optimization pipeline, we conduct an ablation study by training our model on the optimized dataset, replacing the standard LLaVA-OV 1.5 full midtrain data. Performance is measured across nine standard multimodal benchmarks. As summarized in Table 8, the optimized caption pipeline leads to consistent improvements across all evaluated tasks.

The evaluation results show consistent performance gains across multiple capabilities:

1. **Text and Document Comprehension:** The most notable gains are observed on document- and

text-heavy benchmarks, such as InfoVQA (+5.62), OCRBench (+3.80), ChartQA (+1.52), and DocVQA (+1.08). This aligns with the explicit OCR quality metric incorporated in the evaluation sub-agent, which encourages the prompt optimization process to retain dense visual text details in the captions.

2. **Generalization Capabilities:** Performance gains also extend to general visual reasoning and perception benchmarks, including RealWorldQA (+3.79) and MMBench-CN (+1.37). This indicates that multi-aspect feedback helps improve caption quality globally without leading to narrow task-specific overfitting.

Finding 6: **AI-Driven Data Pipeline Optimization Enhances Downstream Performance.** Iterative dataset optimization guided by multi-dimensional LLM rubric scoring consistently yields global quality gains across diverse benchmarks. When linguistic prompt tuning hits a ceiling, the agent further extends to harness-level code modifications—such as video timestamp overlays—enabling effective code-prompt co-design.

Emergent Behavior in Video Caption Pipeline. During the iteration of the video captioning pipeline, the evaluation sub-agent identified a recurring error pattern: the caption model GPT-4.1 frequently struggled with precise temporal localization and timestamp grounding when relying solely on generic text prompts. Modifying the prompt text alone proved insufficient to resolve this bottleneck due to the challenge of implicit temporal alignment across video frames. To address this limitation, the pipeline established a synergy between harness code and prompted model skills. Instead of requiring the model to infer time implicitly through text instructions, the pipeline harness code modifies the input by drawing sequential timestamp overlays directly onto the sampled video frames prior to model encoding. The prompt (skill) then guides the model to utilize its visual OCR recognition capabilities to explicitly anchor events to these rendered timestamps. By converting implicit temporal tracking into a visual text-grounding task, this code-prompt co-design helps reduce temporal hallucinations and improves caption consistency across extended video sequences.

AI-Guided Training Diagnostics & Recipe Optimization. Through iterative AI-driven diagnostics, we systematically identified critical data and architectural bottlenecks, which directly guided our training configurations. First, the evaluation agent diagnosed a distinct vulnerability in the model's precise temporal localization capabilities; to remedy this, we strategically supplemented our dataset with high-quality temporal grounding and localization data. Second, scaling investigations revealed that training with a larger number of frames consistently boosts long-video performance, whereas incorporating dedicated long-video SFT data yielded no noticeable improvements, and scaling the spatial resolution beyond 384 pixels offered only marginal gains. Consequently, we adopted a spatial resolution of 384 and set the temporal length of Stage 3 to 384 frames. Finally, we discovered that increasing the RoPE base frequency (θ) to 8M was highly beneficial to stabilizing and enhancing long-context sequence modeling over extended timelines. Together, these AI-derived diagnostics uniquely determined our final training length, hyperparameters, and data recipes.

6.2 Zero-Vision SFT: Unlocking Agentic Capabilities

Motivation: Rethinking Visual Post-Training. Despite its strong architectural foundation, our primary model, Mage-VL, currently exhibits a capability gap on complex agentic tasks compared to Qwen3-VL. This gap largely stems from computational resource constraints, the lack of joint text-multimodal mixed pre-training, and omitted RL post-training. To address these limitations

Table 9 Effectiveness of Skipping Vision SFT before Multimodal RL. We compare our proposed *Zero-Vision SFT + RL* pipeline against the standard *OV1.5 Quick Start + RL* baseline across 24 multimodal benchmarks. Δ denotes the net improvement of Zero-Vision RL over Quick Start RL ((B) − (C)). Bold text indicates superior performance between the two post-RL models.

	Zero-Vision Paradigm		Baseline Paradigm	
Benchmark	**(A) SFT Only**	**(B) + RL (Ours)**	**(C) QS + RL**	**Δ (B vs. C)**
General VQA				
MMStar [140]	48.73	**54.80**	50.33	+4.47
MMBench-EN [138]	81.10	**83.69**	81.25	+2.44
MMBench-CN [138]	79.16	81.59	**82.91**	−1.32
MME-RealWorld-CN [145]	34.32	43.40	**43.82**	−0.42
MME-RealWorld-EN [145]	45.25	47.55	**49.31**	−1.76
SeedBench-Image [142]	75.13	**75.86**	73.78	+2.08
SeedBench-2-Plus [143]	59.95	**62.85**	60.25	+2.60
CV-Bench [10]	68.76	**73.88**	69.14	+4.74
RealWorldQA [139]	59.08	55.69	**62.74**	−7.05
Category Avg.	61.28	**64.37**	63.73	+0.64
Reasoning				
MathVista-Mini [169]	47.10	**53.50**	47.20	+6.30
WeMath-Loose [170]	33.90	**54.86**	42.57	+12.29
MathVision-Mini [171]	22.70	**31.58**	30.59	+0.99
MathVision-Test [171]	25.13	**38.26**	31.22	+7.04
MathVerse [172]	34.31	**45.48**	33.47	+12.01
MMMU-Val [173]	47.11	**54.22**	47.00	+7.22
DynaMath-Worst [174]	16.37	**22.36**	14.37	+7.99
MMMU-Pro-Standard [175]	32.43	**39.25**	31.39	+7.86
MMMU-Pro-Vision [175]	25.72	**32.49**	22.95	+9.54
Category Avg.	31.64	**41.33**	33.42	+7.92
OCR & Chart				
CharXiv-Description [137]	72.13	**74.48**	47.18	+27.30
CharXiv-Reason [137]	25.30	**29.20**	23.80	+5.40
ChartQA [129]	60.92	**72.84**	71.08	+1.76
OCRBench [130]	**63.40**	61.20	57.60	+3.60
Category Avg.	55.44	**59.43**	49.92	+9.52
Others				
AI2D [128]	65.90	71.99	**73.84**	−1.85
PixmoCount [56]	38.01	**41.76**	27.15	+14.61
Category Avg.	51.96	**56.88**	50.50	+6.38
Overall Average	48.41	**54.28**	48.96	+5.33
Win Count (vs. QS RL)	–	**19 / 24**	5 / 24	–
Optimal RL Steps	–	**2,175**	4,415	*-50.7% steps*

under tight compute budgets, two prevailing assumptions often act as bottlenecks: (1) explicit visual-language SFT (Image SFT) is universally considered indispensable before post-training, and (2) Reinforcement Learning (RL) is widely believed to be ineffective on small-data VLM regimes (such as the LLaVA-OV scale). Contrary to these common beliefs, we hypothesize that explicit visual SFT during the instruction-tuning phase may actually induce catastrophic forgetting of intrinsic textual reasoning capabilities, thereby creating a performance ceiling for subsequent RL exploration. Drawing inspiration from Kimi-2.5 [57], we explore whether replacing visual SFT with high-quality, pure-text reasoning instruction tuning post-midtraining can unlock potent multimodal RL capabilities.

Experimental Pipeline and Fine-Grained Empirical Results. We validate our hypothesis using the open-source **LLaVA-OV-1.5 Quick Start dataset**. Training all models from scratch, we compare two distinct pipelines. The standard baseline follows a three-stage workflow: foundational midtraining on LLaVA-OV-1.5 Quick Start data (Stage 1), standard visual instruction tuning (Stage 2 SFT), and OpenMMReasoner RL [176] (Stage 3). In contrast, our proposed *Zero-Vision SFT (+ RL)* strategy

bypasses visual SFT entirely: in Stage 1, we combine LLaVA-OV-1.5 Quick Start midtraining with pure-text math and code instruction data from `nvidia/Nemotron-Pretraining-SFT-v1` [177] under an equivalent token budget to standard visual SFT, before directly applying OpenMMReasoner RL (Stage 2 RL) without any visual instruction tuning.

As detailed in Table 9, replacing visual SFT with pure-text Zero-Vision SFT in Stage 1 prior to multimodal RL yields impressive performance gains across 24 multimodal image benchmarks, achieving an overall average accuracy of **54.28%** vs. **48.96%** (+5.33% absolute margin) and winning in **19 out of 24** tasks. Crucially, Zero-Vision RL reaches optimal convergence at only **2,175 steps**, reducing required RL iterations by over **50.7%** compared to the Quick Start baseline (4,415 steps).

Analyzing category-level metrics demonstrates that preserving pure-text reasoning capabilities pays off significantly. In **Reasoning** tasks (+7.92% avg. gain), Zero-Vision RL surges across all 9 benchmarks, most notably on WeMath-Loose (**54.86%** vs. 42.57%, +12.29%), MathVerse (**45.48%** vs. 33.47%, +12.01%), and MMMU-Pro-Vision (**32.49%** vs. 22.95%, +9.54%). For **OCR & Chart** understanding (+9.52% avg. gain), direct multimodal RL effectively aligns textual logic with dense visual structures, yielding massive jumps on CharXiv-Desc (**74.48%** vs. 47.18%, +27.30%) and ChartQA (**72.84%** vs. 71.08%). Even in **General VQA** (+0.64% avg. gain) and **Others** (+6.38% avg. gain), direct RL grounds perceptual features effectively without visual SFT, achieving noticeable improvements on MMStar (**54.80%** vs. 50.33%), CV-Bench (**73.88%** vs. 69.14%), and counting tasks like PixmoCount (**41.76%** vs. 27.15%, +14.61%).

Finding 7: Bypassing Visual SFT Unlocks Multimodal RL and Bridges the Agentic Gap. Challenging the common belief that visual SFT is essential and that RL fails on small-data VLMs, replacing image SFT with high-quality text SFT unlocks potent multimodal RL capabilities. It provides a compute-efficient technical path to address current agentic limitations.

7 Conclusion and Limitations

We presented Mage-VL, a lightweight 4B-parameter streaming vision–language model centered on its custom-designed ViT-encoder, Mage-ViT. Built upon the principle of codec-aligned sparsity, Mage-ViT is trained from scratch as a codec-agnostic visual front-end that dynamically extracts motion- and entropy-rich patches across traditional (e.g., HEVC) and neural codecs (e.g., DCVC). Integrated with a causal language decoder and an event-driven System 1 gate, Mage-VL provides a unified architecture for offline multimodal reasoning, fine-grained spatial grounding, and proactive streaming interaction.

Empirically, Mage-VL-4B matches flagship baselines on static image understanding while achieving uniform improvements across video QA, temporal grounding, and 2D/3D spatial reasoning benchmarks, accompanied by up to 3.5× wall-clock inference speedups over dense frame sampling. Furthermore, our systematic empirical studies yield seven foundational findings, covering visual pre-training data efficiency, resolution scaling, codec-driven system acceleration, VideoQA SFT redundancy, motion-spatial synergy, AI-driven data pipeline optimization, and Zero-Vision SFT for multimodal RL.

Limitations and Future Work. Despite its strong visual and temporal performance, Mage-VL currently exhibits limitations in complex agentic workflows and mathematical reasoning. These gaps primarily stem from a shortage of high-quality text data in our current training mixture, as well as the omission of RL post-training. In future iterations, we plan to adopt a Zero-Vision SFT alongside RL to bridge these agentic and mathematical capability gaps, while extending Mage-VL to native audio-visual stream fusion and embodied edge applications.

Contributor List

Contributors: Senqiao Yang*, Kaichen Zhang*, Zhaoyang Jia*, Jinghao Guo*, Yifei Shen*†, Xinjie Zhang*†, Xiaoyi Zhang, Haoqing Wang, Xiao Li, Peng Zhang, Xiang An, Yin Xie, Zhening Liu, Xun Guo, Jiahao Li, Shicheng Zheng, Jinglu Wang, Zongyu Guo, Wenxuan Xie, Zihan Zheng, Yuxuan Luo, Bin Li, Yan Lu

*Equal Contribution. †Project Leads (yshenaw@connect.ust.hk, xinjiezhang@microsoft.com).

A Recaptioning System Prompt

This appendix provides the complete system prompt used for large-scale image recaptioning with Qwen3-VL-32B. The original prompt of LLaVA-OV is following:

LLaVA-OneVision captioning system prompt

```
Generate a single coherent, detailed caption for the image that covers all applicable dimensions below.
Please integrate all relevant dimensions into one comprehensive caption rather than addressing each dimension separately.
Ensure the description is visually grounded - every detail should be verifiable from the image itself.


Pre-defined Dimensions:

0. **World Knowledge**:
Provide relevant factual background related to the image content (historical facts, cultural significance, scientific context).
Use specific names, places, and concepts rather than vague references.
Example: "The Eiffel Tower" instead of "a famous landmark".

1. **Character Name**:
If identifiable characters appear from movies, TV shows, anime, comics, literature, games, or virtual idols, state their names.

2. **Scene Description**:
Provide an overview of the image, identifying key objects, people, and interactions.
Classify and describe each element with specific attributes: size, color, material, texture.

3. **Actions and Interactions**:
Describe actions taking place and how people/objects interact.
Detail dynamic elements (e.g., running, jumping, flying, waving) and their state.

4. **Context and Environment**:
Describe the setting: location (indoor/outdoor), time of day, weather, background elements.
Explain how the environment contributes to the overall scene and mood.

5. **Emotion and Sentiment**:
If people are present, describe emotional states based on body language and facial expressions.
What mood or tone does the image convey (e.g., happiness, sadness, tension, peace)?

6. **Relationships and Spatial Arrangement**:
Explain how objects and people are positioned relative to one another (next to, above, to the right of).
Consider foreground, background, and overall spatial composition.

7. **Color and Texture**:
Describe the color palette and texture details (smooth, rough, soft).
How do color and texture contribute to the atmosphere or style?

8. **Symbolism or Abstract Interpretation**:
If relevant, interpret symbolic or abstract elements.
What deeper meanings or metaphors can be inferred? How do they relate to broader themes?

9. **Lighting and Shadows**:
Observe lighting conditions (sunlight, artificial light) and shadow/reflection patterns.
Note light intensity and its contribution to mood or focal points.

10. **Details and Fine Elements**:
Focus on intricate details (wrinkles in clothing, surface textures, distinct features).
These elements may carry significant meaning or provide vivid precision.

11. **Perspective and Composition**:
Describe viewpoint (aerial, eye-level, side view) and composition (symmetry, balance, focal point).
How does perspective affect viewer perception?

12. **Time and Season**:
Infer time of day or season from visual cues (light quality, weather, clothing).
Examples: winter snow scene, summer beach, autumn forest.

13. **Target Audience**:
Consider if the image suggests a specific target audience or purpose.
Adjust complexity and terminology accordingly (e.g., technical vs. general audience).

14. **OCR (Text in Image)**:
If text appears, transcribe it in the original language and provide English translation in parentheses.
Example: "书本 (book)". Explain the meaning within context.

15. **Person Description**:
Describe physical features (age, gender, hairstyle, clothing), movements, and expressions.
Explain relationship to the surrounding environment.
Use singular pronouns (he/she) for single individuals rather than "they".

16. **Mathematics**:
Describe mathematical concepts: geometric shapes, equations, numeric values, relationships.
For graphs, describe axes, scales, and key points.
Explain how mathematical operations are visualized.

17. **Information Extraction**:
Extract textual and contextual information.
For documents, transcribe content accurately.
For GUI or structured data, describe layout, labels, and functionality.

18. **Planning**:
Identify sequences or logical arrangements.
```

```
For processes, explain steps in correct order.
For puzzles or games, provide rules and possible solutions.

19. **Coding**:
If code appears, output it line by line first.
Explain syntax, function, and purpose.
For debugging tasks, identify issues or provide equivalent code in other languages.

20. **Perception**:
Provide detailed perception-based descriptions.
Identify object attributes (color, shape, size) and spatial relationships.
For facial analysis or pose estimation, include expressions, poses, and physical traits.

21. **Metrics**:
Evaluate image quality, authenticity, and adherence to caption content.
For comparative tasks, provide constructive feedback or preference reasoning.

22. **Science**:
Explain scientific content or phenomena depicted.
Include details on experiments, natural phenomena, or theoretical concepts with relevant terminology.


Output Requirements:
- Generate a single, coherent caption integrating all applicable dimensions
- Do NOT format as separate sections (e.g., **Scene Description**, **Actions**, etc.)
- Omit dimensions that are not applicable to the image
- Maintain a detailed, comprehensive description
- Ensure all details are visually verifiable from the image
```

The prompt is obtained through the iterative optimization procedure described in Section 4.2.1, with the goal of improving visual coverage, reducing redundant descriptions, maintaining coherent organization, and preserving rendered text with high OCR fidelity. The final prompt is used to generate the approximately 350M image–caption pairs employed in our image-caption training stage.

Optimized image captioning system prompt

```
Write a single, detailed, and strictly objective caption for this image. Describe only what is directly visible. Be assertive and precise
      --state facts, not possibilities.

Writing style: Write in natural, flowing prose as a single coherent paragraph. Integrate details organically into sentences rather than
      listing attributes one by one. Use varied sentence structures--combine related details within the same sentence using clauses,
      appositives, and conjunctions. Transition smoothly between subjects (e.g., from foreground to background, from one object to the
      next) so the description reads like a cohesive narrative, not a feature inventory.

BAD (mechanical listing): "The dog has brown fur. The dog has floppy ears. The dog's eyes are dark. The dog is sitting. The dog is on a
      wooden floor."
GOOD (flowing prose): "A brown dog with floppy ears and dark eyes sits on a wooden floor, its front paws resting side by side."

BAD (repetitive structure): "The shirt is blue. The pants are black. The shoes are white. The hat is red."
GOOD (integrated): "The person wears a blue shirt tucked into black pants, paired with white shoes and a red hat."

Cover the following aspects where applicable (do NOT use headers, bullet points, or labels):

- **Objects & Attributes**: Key objects, people, and elements with observable attributes: color, shape, size, material, texture, quantity.

- **Fine Details**: Intricate or small-scale details that add precision--wrinkles in fabric, scratches on surfaces, visible stitching,
      watermarks, tags, labels, logos, serial numbers, or distinctive marks.
- **Text & Data (OCR)**: Transcribe any visible text exactly as it appears in its original language. If the text is not in English,
      provide an English translation in parentheses immediately after (e.g., "Libro (book)", "Sortie (Exit)"). Briefly explain the
      meaning or role of the text within the context of the image (e.g., a sign, a label, a title, a watermark). Reproduce key data from
      tables, charts, and labels accurately, including column headers, row labels, and numerical values.
- **Documents & GUI**: For screenshots, documents, or structured interfaces, describe the layout, navigation elements, buttons, menus,
      input fields, and content hierarchy. Transcribe visible text content accurately.
- **Spatial Layout**: Positions, arrangements, and spatial relationships between elements. State each spatial relationship only once.
- **Actions & Poses**: Observable body positions and actions factually (e.g., "mouth open, eyes wide" not "expressing surprise"). Never
      infer emotions, intentions, or mental states.
- **People**: Describe age range, gender, hairstyle, clothing, posture, and physical features factually. For facial expressions, state
      only the physical configuration (e.g., "corners of the mouth turned upward") without labeling emotions. Use singular pronouns (he/
      she) for single individuals rather than "they."
- **Setting & Environment**: Indoor/outdoor, background elements, weather, time of day, and season only when indicated by visible cues (e.
      g., snow on ground -> winter; long shadows and orange sky -> late afternoon).
- **Color Palette & Lighting**: Colors, lighting direction/type (sunlight, artificial, diffused), shadow patterns, and reflections
      factually.
- **Perspective & Composition**: Viewpoint (top-down, eye-level, side view, aerial) and compositional structure (symmetry, focal point,
      leading lines, depth of field).
- **Identifiable Entities**: Use specific names for recognizable people, places, brands, characters, or landmarks. Always prefer precise
      identification over generic descriptions.
  Examples: "The Eiffel Tower" instead of "a tall metal tower"; "Pikachu" instead of "a yellow cartoon creature"; "a Nike swoosh logo"
      instead of "a checkmark-shaped logo"; "Barack Obama" instead of "a man in a suit."
  This applies to fictional characters from anime, comics, movies, TV shows, games, and virtual idols--state their names when
      recognizable.
- **Sequential & Multi-panel Content**: For images showing step-by-step processes, numbered panels, comic strips, or before/after
      comparisons, describe each step or panel in order, noting panel numbers or labels if present.
- **Code & Programming**: If code is visible, transcribe it line by line. Describe the programming language, syntax structure, and
      visible function/variable names. Note any syntax highlighting colors.
```

```
- **Mathematics & Graphs**: For equations, render them in LaTeX. For graphs and charts, describe axes (labels, units, scale), data points
      , trends, and legends. For geometric figures, describe shapes, angles, and measurements.
- **Scientific Content**: For scientific images (experiments, specimens, diagrams, phenomena), use appropriate technical terminology.
      Describe apparatus, measurements, labels, and observable processes.

Strict rules--violation of any rule makes the caption unacceptable:

1. NEVER start with "The image" in any form (e.g., "The image shows/displays/features/presents/depicts/is/contains"). Instead, begin
      directly with the subject: "A dog...", "Two women...", "An aerial view of...", "A close-up of...".
2. NEVER use any form of these words anywhere: "suggest/suggesting/suggests", "indicate/indicating/indicates", "imply/implying/implies",
      "evoke/evoking/evokes", "convey/conveying/conveys", "hint/hinting/hints at", "allude/alluding/alludes". State the visual fact
      directly without inferring purpose or meaning.
   BAD: "green leaves suggesting spring" -> GOOD: "green leaves"
   BAD: "open mouth indicating speech" -> GOOD: "open mouth"
   BAD: "a pattern suggesting movement" -> GOOD: "a curved, flowing pattern"
   BAD: "shadows suggesting depth" -> GOOD: "shadows fall across the surface"
   BAD: "colors hinting at sunset" -> GOOD: "orange and pink tones in the sky"
3. NEVER use "creating", "creates", or "created" to describe visual effects or impressions. Use concrete verbs instead: casts, produces,
      forms, results in, yields, adds, gives, lends.
   BAD: "red light creating a warm tone" -> GOOD: "red light casts a warm-toned glow"
   BAD: "lines creating depth" -> GOOD: "lines converge toward a vanishing point"
   BAD: "contrast creating emphasis" -> GOOD: "the high contrast between the dark background and bright subject draws focus"
   BAD: "bokeh creating a soft background" -> GOOD: "bokeh dots fill the out-of-focus background"
   BAD: "shadows creating texture" -> GOOD: "shadows accentuate the surface texture"
4. NEVER use "appears to", "appears", "seems to", "seems", "likely", "possibly", "presumably", "probably", or "perhaps"--state what IS
      visible as fact.
   BAD: "The surface appears rough" -> GOOD: "The surface is rough"
   BAD: "appears to be a cat" -> GOOD: "a cat"
5. NEVER use "overall" anywhere in the caption. Do not summarize the scene at the end; instead, continue describing visible details.
6. NEVER add emotional, atmospheric, or aesthetic commentary. Banned words include: "warm" (except for color temperature like "warm-toned
       lighting"), "serene", "calm", "peaceful", "gentle", "intimate", "heartfelt", "elegant", "cozy", "dramatic", "striking", "vibrant
      mood", "whimsical", "enchanting", "playful", "inviting", "charming", "delicate", "lively", "cheerful", "magical", "dynamic" (except
       physics), "graceful", "sense of", "gives a feeling of", "atmosphere", "mood".
7. NEVER describe what is absent. Only describe what IS present and visible. Banned phrases: "no visible", "no text", "no people", "no
      other", "no additional", "no distinct", "no further", "nothing else", "without any", "lacks". Simply omit anything not present.
   BAD: "No text is visible. No people are present. No other objects are seen."
   BAD: "no additional decorations or objects in the visible area"
   BAD: "no distinct shadows are cast"
   GOOD: (simply omit--say nothing about absent elements)
8. NEVER summarize or restate information already described. Each fact appears exactly once.
9. Every statement must be visually verifiable from the image alone.
10. Omit any aspect that does not apply.
11. Do NOT write in a mechanical, checklist style. Avoid sequences of short, isolated sentences that each describe one attribute. Instead
      , merge related details into richer, compound sentences.

Before finalizing, perform this mandatory self-check word by word:
- Scan for: "suggesting", "indicating", "implying", "hinting", "creating", "creates" -> replace with concrete verbs
- Scan for: "appears", "seems", "likely", "possibly", "overall" -> delete or rewrite as fact
- Scan for: "no visible", "no other", "no additional", "no distinct" -> delete the entire sentence
- Scan for: "warm", "gentle", "calm", "delicate", "striking", "elegant" -> delete the adjective
- Check the first word: if "The image" -> rewrite to start with the subject
- Check for 3+ consecutive short sentences -> combine them

Examples:

Example 1 (Table image):
A table titled "Table 2. Variation in content of lipids, EPA and DHA in livers from cod fish during 21 months" spans the full width of
      the frame against a white background, with the title set in bold black serif text centered above the grid. The header row is filled
       in solid red with three column labels in white bold text--"Total lipid content (g/100g)", "EPA (% of total fatty acids)", and "DHA
       (% of total fatty acids)"--separated by thin black vertical borders. Five data rows sit below, each beginning with a fish species
      name in the leftmost column: Ling, Saithe, Hake, Cod, and Monkfish, followed by their corresponding numerical ranges--Ling: 51-66,
      5.1-7.7, 10.2-12.9; Saithe: 48-76, 4.4-9.0, 10.5-13.3; Hake: 42-67, 5.5-7.7, 11.7-13.0; Cod: 50-64, 6.0-8.6, 12.3-14.4; Monkfish:
      31-43, 4.6-6.3, 10.8-13.1. All numerical values are center-aligned in black regular-weight serif font, and thin black borders
      outline every cell. The top-down perspective captures the complete table with uniform row heights and consistent cell spacing.

Example 2 (Outdoor scene with people):
A crowded outdoor market extends along a narrow cobblestone street, with vendor stalls on both sides sheltered under canvas tarps in
      faded blue, green, and beige. In the foreground, a woman wearing a long olive-green coat and brown leather boots reaches toward a
      pile of red and yellow apples arranged in tiered wooden crates, while the vendor--a man in a gray wool cap and dark blue apron over
       a plaid flannel shirt--extends a small brown paper bag toward her with his right hand. To her left, a child in a red puffy jacket
      and jeans holds a half-eaten pretzel in one hand and grips the woman's coat with the other, standing at hip height beside her.
      Further down the street, clusters of shoppers move between stalls stacked with root vegetables, hanging sausages, and loaves of
      bread on wire racks, their forms becoming progressively blurred with distance. Strings of bare Edison-style light bulbs hang in
      parallel rows above the stalls at roughly three-meter intervals, and late-afternoon sunlight enters from the upper right, casting
      long shadows to the left across the cobblestones and illuminating the stone facades of three- and four-story residential buildings
      lining both sides of the street. The eye-level perspective is centered in the street, with the converging lines of stalls and
      overhead lights leading toward a clock tower with a copper-green spire at the far end.

Example 3 (Animal close-up):
A chimpanzee with dense black fur sits on a thick horizontal tree branch, its left hand gripping the rough bark while its right hand
      holds a partially peeled banana near its open mouth, the lower lip pulled down to reveal a row of flat white teeth. The chimp's
      dark gray face is creased with deep wrinkles around the eyes, and it has a broad, flat nose with flared nostrils, small rounded
      ears set high on the head and partially obscured by tufts of coarse fur, and sparse white hairs along its chin and lower jaw. Its
      dark brown eyes, with visible round pupils, are directed downward toward the banana. Behind the chimpanzee, layers of out-of-focus
      green foliage--broad tropical leaves, thin vertical stems, and patches of brown bark from adjacent trees--fill the frame in varying
       shades of emerald and olive. A thin shaft of sunlight cuts diagonally from the upper left through the canopy, catching the fur
      along the chimp's right shoulder in a bright highlight while the rest of the body remains in dappled shade. The eye-level
      perspective places the chimpanzee slightly left of center, with the branch running horizontally across the lower third of the frame
       and the dense canopy occupying the upper two-thirds.

Now write the caption for the provided image:
```

Video captioning system prompt

```
# Video Description Prompt & Requirements

Please describe the video in detail, including but not limited to the user pre-defined dimensions. Please make sure your description is
      visually grounded for the user-provided video, namely the user can find visual cues in the video for your generated video caption.

> **IMPORTANT:** Your response must be in English plain text format only. Do not use any markdown formatting such as ‘**bold**‘, ‘*italic
      *‘, ‘### headers‘, or any other markdown symbols. Use only regular text without any special formatting characters.

---

## User Pre-defined Dimensions

### 0. Context and Environment
Describe the setting of the image, including:
* **Location:** Indoor or outdoor
* **Time & Weather:** Time of day, weather conditions
* **Background Elements:** Sky, buildings, roads, etc.
* **Impact:** How the environment contributes to the overall scene and whether the setting enhances the mood or theme.

### 1. Main Subject of the Video
Identify the primary subject of the video (e.g., a male athlete, a running lion, a plane taking off, a flying bird, a swimming fish, a
      female dancer performing a dance).
* *Note:* If the main subject is a character, you **must** specify his or her name (including characters from movies, TV shows, anime,
      comics, literature, video games, virtual idols, or virtual characters).

### 2. Actions and Interactions
Describe any named actions occurring in the video (e.g., butterfly stroke, pole vault, jazz dance). Specify who is performing these
      actions and how they interact with other objects or people.

### 3. Motion Detail Description
For dynamic elements in the video:
* **State:** Running, jumping, flying, waving, swimming, cooking, etc.
* **Direction of Motion:** Left to right, front to back, clockwise/counterclockwise rotation, etc.
* **Speed of Movement:** Sudden acceleration, gradual deceleration, etc.
* **Sequence of Actions:** Step-by-step chronological order (e.g., *first open the refrigerator, then take out the food, next wash the
      food, and finally cook it*).

### 4. Background Changes
If the environment or scene in the video changes, provide a detailed description of how the background evolves over time.
* *Example 1:* At around the 30-second mark in the latter half of the video, the weather shifts from sunny to rainy.
* *Example 2:* Around the 20-second mark in the middle of the video, the street gradually transitions from noisy and crowded to quiet and
       empty.
* *Example 3:* At approximately 50 seconds, gentle ripples begin to appear on an otherwise calm sea surface, escalating by 80 seconds
      into turbulent, stormy waves.

### 5. Highlight Moments
If the video contains significant or impactful events (e.g., a soccer goal, a person falling, a tiger hunting prey, a boxer knocking down
       an opponent), specify the exact time each event occurs and provide a detailed description.
* *Example 1:* At approximately 15 seconds, the player wearing a white jersey, number 5, kicks the soccer ball into the goal.
* *Example 2:* At around 3 seconds, the boxer in red shorts knocks down his opponent, who is dressed in black.

### 6. First-Person Perspective
If the video is shot from a first-person point of view (POV):
* Clearly state the main activity being performed (e.g., cooking, painting, playing soccer).
* Provide a detailed chronological description of the subject’s actions and their timing (e.g., *at around 2 seconds, the viewer opens a
      laptop; at 5 seconds, they navigate to a video website; at 8 seconds, they open a can of beverage*).

---

## Response Generation Rules

* **Output Style:** Do **not** generate your response split by dimension (e.g., do not write ‘**Context and Environment**: ...‘). Provide
       **one single overall image/video caption** that naturally integrates all dimensions into a cohesive text.

### Critical Formatting Requirements

1. **Structure:** The format of your answer should be a **single continuous paragraph**.
2. **Plain Text Only:** Use **ONLY plain text** -- strictly NO markdown formatting whatsoever (no ‘**‘, ‘*‘, ‘###‘, etc.).
3. **No Special Symbols:** Do not use any special characters for emphasis or formatting.
4. **Tone & Flow:** Write in natural, flowing sentences without any markup symbols.
```

References

[1] Anthropic. Claude fable 5 & claude mythos 5 system card. Technical report, Anthropic, 6 2026. URL https://www.anthropic.com/system-cards.

[2] Google DeepMind. Gemini 3 pro model card. Technical report, Google DeepMind, 11 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

[3] OpenAI. Gpt-5.5 system card. Technical report, OpenAI, 4 2026. URL https://openai.com/index/gpt-5-5-system-card/.

[4] Hans Moravec. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, 1988.

[5] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv:2502.13923*, 2025.

[7] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, Dengnian Chen, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv:2504.10479*, 2025.

[8] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision-1.5: A family of fully open vision–language models with the llava-onevision-1.5 mid-training and instruct data. *arXiv preprint*, 2025.

[9] Biao Yang, Bin Wen, Boyang Ding, et al. Kwai keye-vl 1.5 technical report. *arXiv:2509.01563*, 2025.

[10] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *NeurIPS*, 2024.

[11] Kimi Team. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[12] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa,

et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[13] Peter Sterling and Simon Laughlin. *Principles of neural design*. MIT press, 2015.

[14] Jeremy E Niven and Simon B Laughlin. Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology*, 211(11):1792–1804, 2008.

[15] Tim Gollisch and Markus Meister. Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron*, 65(2):150–164, 2010.

[16] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[17] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[18] Feilong Tang, Xiang An, Yunyao Yan, Yin Xie, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Chunyuan Li, Shikun Feng, Changrui Chen, Huajie Tan, Ming Hu, Manyuan Zhang, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. OneVision-Encoder: Codec-aligned sparsity as a foundational principle for multimodal intelligence. *arXiv preprint arXiv:2602.08683*, 2026.

[19] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression (DCVC). *NeurIPS*, 2021.

[20] Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Chunyuan Li, Bo Li, Ziwei Liu, Jiankang Deng, et al. LLaVA-OneVision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

[21] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

[22] Jyoti Aneja, Michael Harrison, Neel Joshi, Tyler LaBonte, John Langford, and Eduardo Salinas. Phi-4-reasoning-vision-15b technical report, 2026. URL https://arxiv.org/abs/2603.03975.

[23] James R Bergen and Edward H Adelson. The plenoptic function and the elements of early vision. *Computational models of visual processing*, 1(8):3, 1991.

[24] Yifei Shen, Bo Li, and Xinjie Zhang. Skillopt-lite: Better and faster agent self-evolution via one line of vibe. *arXiv preprint arXiv:2607.03451*, 2026.

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ICML*, 2021.

[27] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *ICCV*, 2023.

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

[29] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 2021.

[30] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. AdaViT: Adaptive vision transformers for efficient image recognition. In *CVPR*, 2022.

[31] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *ICLR*, 2023.

[32] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision (ECCV)*, 2024.

[33] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[34] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning (ICML)*, 2021.

[35] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[36] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[37] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[38] Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023.

[39] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[40] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. SlowFast-LLaVA: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024.

[41] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[42] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

[43] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Chen, Yinan Zhu, Haian Sun, Yali He, Yu Wang, and Limin Wang. VideoChat-Flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2025.

[44] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Compressed video action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[45] Yang Jin, Zhicheng Sun, Kun Xu, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, Kun Gai, and Yadong Mu. Video-LaVIT: Unified video-language pre-training with decoupled visual-motional tokenization. *ICML*, 2024.

[46] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[47] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023.

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

[49] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[50] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[51] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[52] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[53] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CVPR*, 2024.

[54] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. DeepSeek-VL: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[55] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12B. *arXiv preprint arXiv:2410.07073*, 2024.

[56] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. *arXiv preprint arXiv:2409.17146*, 2024.

[57] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[58] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[59] Chaorui Deng et al. BAGEL: A unified multimodal foundation model for image understanding, generation, and editing. *arXiv preprint*, 2025.

[60] Yuying Cui et al. Emu3.5: Native multimodal models. *arXiv preprint*, 2025.

[61] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025.

[62] Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025.

[63] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[64] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[65] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[66] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[67] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[68] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[69] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. VideoLLM-online: Online video large language model for streaming video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[70] Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. VideoLLM knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format. *arXiv preprint arXiv:2411.17991*, 2024.

[71] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. *arXiv preprint arXiv:2501.03218*, 2025.

[72] Xin Ding, Hao Wu, Yifan Yang, Shiqi Jiang, Qianxi Zhang, Donglin Bai, Zhibo Chen, and Ting Cao. Streammind: Unlocking full frame rate streaming video dialogue through event-gated cognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13448–13459, 2025.

[73] Video Understanding Team of JoyAI-VL @ Joy Future Academy, JD. Joyai-vl-interaction: Real-time vision-language interaction intelligence. Technical report, Joy Future Academy, JD, June 2026.

[74] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-VStream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024.

[75] Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams, 2026. URL https://arxiv.org/abs/2510.09608.

[76] Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, et al. InternLM-XComposer2.5-OmniLive: A comprehensive multimodal system for long-term streaming video and audio interactions. *arXiv preprint arXiv:2412.09596*, 2024.

[77] Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Livecc: Learning video llm with streaming speech transcription at scale, 2025. URL https://arxiv.org/abs/2504.16030.

[78] Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. Matchtime: Towards automatic soccer game commentary generation, 2024. URL https://arxiv.org/abs/2406.18530.

[79] Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the gap for mllms to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024.

[80] Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. OVO-Bench: How far is your video-llms from real-world online video understanding? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[81] Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (HEVC) standard. In *IEEE Transactions on Circuits and Systems for Video Technology*, 2012.

[82] Zhaoyang Jia, Bin Li, Jiahao Li, Wenxuan Xie, Linfeng Qi, Houqiang Li, and Yan Lu. Towards practical real-time neural video compression. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12543–12552, 2025.

[83] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

[84] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

[85] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400m: Open dataset of CLIP-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[86] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. COYO-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

[87] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: An open web-scale filtered dataset of interleaved image-text documents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[88] Chunyu Xie, Heng Cai, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Jianfei Song, Henrique Morimitsu, Lin Yao, Dexin Wang, Xiangzheng Zhang, Dawei Leng, Baochang Zhang, Xiangyang Ji, and Yafeng Deng. CCMB: A large-scale chinese cross-modal benchmark. *arXiv preprint arXiv:2205.03860*, 2023.

[89] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[90] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[91] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[92] Qwen Team. Qwen3 technical report. *arXiv preprint*, 2025.

[93] Xinjie Zhang, Peng Zhang, Shicheng Zheng, Jinghao Guo, Zhaoyang Jia, Yifei Shen, Xun Guo, Yuxuan Luo, Jiahao Li, Wenxuan Xie, Fanyi Pu, Xiaoyi Zhang, Kaichen Zhang, Zongyu Guo, Tianci Bi, Dongnan Gui, Zhening Liu, Zimo Wen, Zihan Zheng, Senqiao Yang, Xiao Li, Jinglu Wang, Bin Li, and Yan Lu. Mage-flow: An efficient native-resolution foundation model for image generation and editing. *arXiv preprint arXiv:2607.19064*, 2026.

[94] Microsoft Mage Team. Ai4ai at scale: A full-pipeline system for enhancing llm agentic capabilities. Technical report, Microsoft, 2026.

[95] Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. FineVision: Open data is all you need. *arXiv preprint arXiv:2510.17269*, 2025.

[96] Yi Zhang, Bolin Ni, Xin-Sheng Chen, Heng-Rui Zhang, Yongming Rao, Houwen Peng, Qinglin Lu, Han Hu, Meng-Hao Guo, and Shi-Min Hu. Bee: A high-quality corpus and full-stack suite to unlock advanced fully open MLLMs, 2026. URL https://arxiv.org/abs/2510.13795.

[97] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data, 2025. URL https://arxiv.org/abs/2410.02713.

[98] Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. TimeLens: Rethinking video temporal grounding with multimodal LLMs, 2026. URL https://arxiv.org/abs/2512.14698.

[99] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28652–28668, 2026.

[100] Xin Ding, Hao Wu, Yifan Yang, Shiqi Jiang, Donglin Bai, Zhibo Chen, and Ting Cao. Streammind: Unlocking full frame rate streaming video dialogue through event-gated cognition, 2025. URL https://arxiv.org/abs/2503.06220.

[101] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *arXiv preprint arXiv:1311.3618*, 2013.

[102] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[103] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[104] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014.

[105] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

[106] Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen-tau Yih, Shang-Wen Li, and Hu Xu. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025.

[107] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, et al. Multimodal autoregressive pre-training of large vision encoders. *arXiv preprint arXiv:2411.14402*, 2024.

[108] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[109] Kimi Team. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[110] Yingwei Li, Yi Li, and Nuno Vasconcelos. RESOUND: Towards action recognition without representation bias. In *ECCV*, 2018.

[111] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV*, 2011.

[112] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, et al. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS Datasets and Benchmarks Track*, 2023.

[113] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, 2018.

[114] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[115] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, 2025.

[116] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024.

[117] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021.

[118] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, et al. TempCompass: Do video LLMs really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

[119] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[120] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.

[121] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, et al. LVBench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024.

[122] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, et al. MLVU: Benchmarking multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

[123] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *ICCV*, 2017.

[124] Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. *arXiv preprint arXiv:2512.14698*, 2025.

[125] Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025.

[126] Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. DocVQA: A dataset for VQA on document images. In *WACV*, 2021.

[127] Minesh Mathew, Viraj Bagal, Rubèn Pérez-Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. InfographicVQA. *arXiv preprint arXiv:2104.12756*, 2021.

[128] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *arXiv preprint arXiv:1603.07396*, 2016.

[129] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of ACL*, 2022.

[130] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, et al. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 2024.

[131] Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, et al. CC-OCR: A comprehensive and challenging OCR benchmark for evaluating large multimodal models in literacy. *arXiv preprint arXiv:2412.02210*, 2024.

[132] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multi-page DocVQA. *arXiv preprint arXiv:2212.05935*, 2022.

[133] Jordy Van Landeghem, Rubén Tito, Łukasz Borchmann, Michał Pietruszka, et al. Document understanding dataset and evaluation (DUDE). In *ICCV*, 2023.

[134] Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. WebSRC: A dataset for web-based structural reading comprehension. In *EMNLP*, 2021.

[135] Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, et al. ChartQAPro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*, 2025.

[136] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, 2019.

[137] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, et al. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *NeurIPS Datasets and Benchmarks Track*, 2024.

[138] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, et al. MMBench: Is your multi-modal model an all-around player? In *ECCV*, 2024.

[139] xAI. RealWorldQA: A benchmark for real-world spatial understanding. https://x.ai/news/grok-1.5v, 2024.

[140] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, et al. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024.

[141] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[142] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension. In *CVPR*, 2024.

[143] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. SEED-Bench-2-Plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.

[144] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, et al. MMT-Bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *ICML*, 2024.

[145] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, et al. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *ICLR*, 2025.

[146] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[147] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, et al. BLINK: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

[148] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *ACL*, 2024.

[149] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, et al. The all-seeing project V2: Towards general relation comprehension of the open world. In *ECCV*, 2024.

[150] Yipu Wang, Yuheng Ji, Yuyang Liu, Enshen Zhou, Ziqiang Yang, Yuxuan Tian, Ziheng Qin, Yue Liu, Huajie Tan, Cheng Chi, Zhiyuan Ma, Daniel Dajun Zeng, and Xiaolong Zheng. Towards cross-view point correspondence in vision-language models. *arXiv preprint arXiv:2512.04686*, 2025.

[151] Gemini Robotics Team. Gemini robotics: Bringing AI into the physical world. Technical report, Google DeepMind, 2025. arXiv:2503.20020.

[152] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, et al. MMSI-Bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025.

[153] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, et al. SAT: Dynamic spatial aptitude training for multimodal language models. In *COLM*, 2025.

[154] Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

[155] Wentao Ma, Weiming Ren, Yiming Jia, Zhuofeng Li, Ping Nie, Ge Zhang, and Wenhu Chen. VideoEval-Pro: Robust and realistic long video understanding evaluation. *arXiv preprint arXiv:2505.14640*, 2025.

[156] Arushi Goel, Sreyan Ghosh, Vatsal Agarwal, Nishit Anand, Kaousheik Jayakumar, Lasha Koroshinadze, Yao Xu, Katie Lyons, James Case, Karan Sapra, et al. Mmou: A massive multi-task omni understanding and reasoning benchmark for long and complex real-world videos. *arXiv preprint arXiv:2603.14145*, 2026.

[157] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, 2018.

[158] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023.

[159] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, et al. VISA: Reasoning video object segmentation via large language models. *arXiv preprint arXiv:2407.11325*, 2024.

[160] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, 2020.

[161] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-Caption: Dense video captioning for soccer broadcasts commentaries. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

[162] Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*, 2026.

[163] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2410.02713.

[164] Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, et al. TimeChat-Online: 80% visual tokens are naturally redundant in streaming videos. *arXiv preprint arXiv:2504.17343*, 2025.

[165] Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, et al. StreamForest: Efficient online video understanding with persistent event memory. *arXiv preprint arXiv:2509.24871*, 2025.

[166] Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. Streaming video instruction tuning. *arXiv preprint arXiv:2512.21334*, 2026.

[167] Gueter Josmy Faure, Jia-Fong Yeh, Min-Hung Chen, Hung-Ting Su, Shang-Hong Lai, and Winston H. Hsu. HERMES: temporal-coherent long-form understanding with episodes and semantics. *arXiv preprint arXiv:2408.17443*, 2025.

[168] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[169] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

[170] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, et al. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

[171] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. *arXiv preprint arXiv:2402.14804*, 2024.

[172] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, et al. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *ECCV*, 2024.

[173] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

[174] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *ICLR*, 2025.

[175] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, et al. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *ACL*, 2025.

[176] Kaichen Zhang, Keming Wu, Zuhao Yang, Bo Li, Kairui Hu, Bin Wang, Xingxuan Li, and Lidong Bing. Openmmreasoner: Pushing the frontiers in multimodal reasoning with an open and general recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19276–19286, 2026.

[177] NVIDIA. Nemotron-Pretraining-SFT-v1 Dataset. https://huggingface.co/datasets/nvidia/Nemotron-Pretraining-SFT-v1, 2025.