# Let RGB Be the Language of Vision

**Timing Yang**[1], **Jinrui Yang**[2], **Xinlong Li**[1*], **Yuhan Wang**[2], **Haoran Li**[3], **Yanqing Liu**[2], **Guoyizhe Wei**[1], **Jixuan Ying**[1*], **Chen Wei**[4], **Rama Chellappa**[1], **Yuyin Zhou**[2], **Cihang Xie**[2], **Alan Yuille**[1], **Feng Wang**[1†]
[1]Johns Hopkins University, [2]UC Santa Cruz, [3]Carnegie Mellon University, [4]Rice University
[†]Project lead, [*]visiting students

This work introduces a unified formulation for vision models, where diverse forms of visual information beyond natural images, such as masks, depth maps, and other structured visual signals, are all represented as RGB images, while general visual tasks can be converted into a common RGB-to-RGB image editing problem. In this paradigm, different types of visual information internally share the same encoding and decoding architecture and parameters as natural images, enabling a single model to transfer across tasks through a unified visual interface, in a way analogous to how language models operate over text. We refer to this formulation as RGB In and RGB Out (**RINO**). Built upon a generic image editing backbone without task-specific fine-tuning, RINO demonstrates robust and competitive zero-shot performance on both dense understanding tasks such as segmentation and depth estimation (where we unify outputs as RGB), and dense-conditioned generation tasks such as pose-to-image generation (where we unify inputs as RGB). We hope this study provides useful insights toward general unified vision-language systems, where diverse visual tasks can be expressed, interpreted, and solved through a shared visual language. Code is available at https://github.com/yangtiming/RINO.

**Date:** July 12th, 2026

## 1 Introduction

In Large Language Models (LLMs), text serves as the universal input and output format for almost all tasks, allowing a well-trained language model to flexibly support diverse downstream applications. In vision, however, this level of unification is still far from being achieved. Different types of visual information typically require their own specific representations, such as RGB for natural images, one-hot masks for segmentation, and continuous geometric maps for depth. This diversity in representation leads to a direct challenge: for each format, we often need to design a dedicated encoder and decoder, together with additional adapter components to connect them. Such barriers in both model structure and information representation make it difficult for vision models to support free-form interactions in the same way as language models, and often limit their ability to effectively generalize beyond individual tasks, leaving most vision foundation models still closer to task-specific experts than truly general-purpose visual learners (Kirillov et al., 2023; Yang et al., 2024a; Rombach et al., 2022; Radford et al., 2021; Oquab et al., 2024).

This naturally raises a question: is there a way for visual information to communicate as freely as text under a unified form? In other words, **can vision develop its own universal interface?** Interestingly, recent studies provide useful insights. Vision Banana (Gabeur et al., 2026) proposes to solve understanding problems such as segmentation, depth estimation, and surface normal estimation with a generative model; for example, instead of predicting a segmentation mask in its conventional format, it generates the mask in RGB space based on the input image. This generative understanding paradigm has also been further verified on open-source image models and shown strong robustness (Liu et al., 2026). Inspired by these observations, we ask whether RGB can be used to unify both the input and output forms of visual information, serving a role similar to text in LLMs and forming a universal "visual language" that humans can naturally interact with.

To this end, we propose **RGB In and RGB Out (RINO)**, a unified visual representation learning paradigm that expresses all visual input and output signals in RGB format. Unlike Vision Banana that evaluates on specific

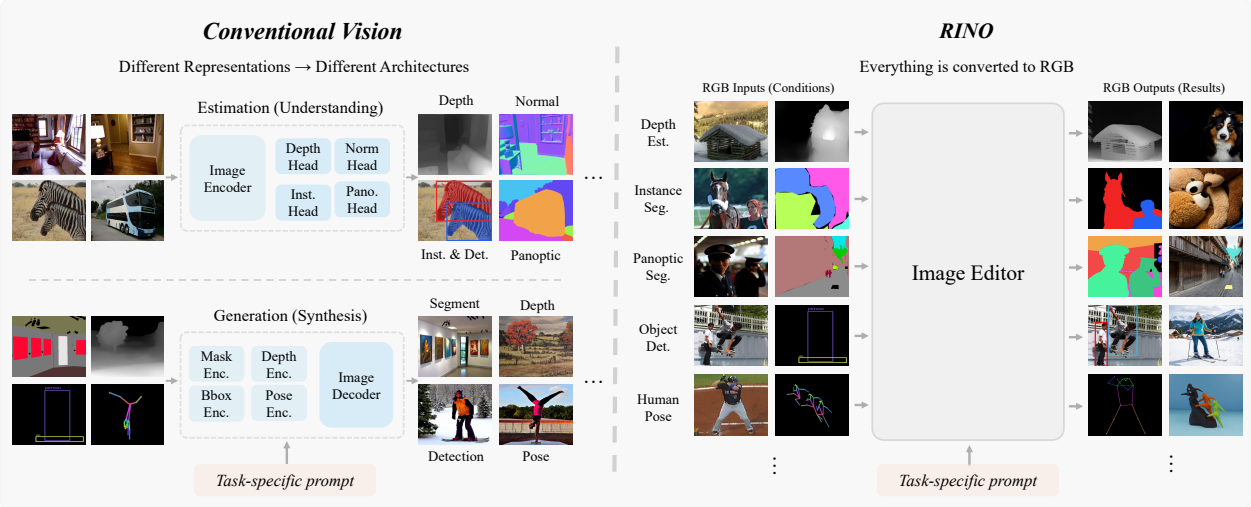


**Figure 1 RINO unifies vision under a single RGB interface.** *Left:* conventional vision ties each task to its own representation and architecture, with task-specific encoders and heads or generators. *Right:* RINO renders every input and output as RGB, so one frozen image editor, driven by a task-specific prompt, handles estimation (depth, normals, segmentation, detection, pose, referring) and the corresponding conditioned generation, all without any task-specific module.

understanding tasks, we aim to investigate how broadly this paradigm can generalize: we evaluate RINO on over 20 vision tasks/benchmarks across multiple domains, including dense understanding, 3D estimation, and conditioned generation, where all visual inputs and outputs are represented in RGB format and share the same encoding/decoding system as natural images. We build RINO upon pretrained image editing models such as Qwen-Image-Edit (Wu et al., 2025) without introducing any auxiliary parameters. For each type of visual information beyond natural images, we only employ lightweight, parameter-free data conversion modules to transform signals between RGB and other formats such as segmentation masks and depth maps, while inside the model, all such visual information is interpreted or generated purely as images.

By constructing this standard RGB/text-to-RGB/text framework, we surprisingly find that RINO can broadly handle all tested understanding and generation tasks in a zero-shot manner. It stably produces visually meaningful results and achieves quantitative performance comparable to in-domain expert models across different applications. For example, on depth estimation, RINO achieves an $\delta_1$ score of 0.938, which approaches Depth Anything (Yang et al., 2024a), a specialist model trained specifically for depth estimation. We further find that for conditioned image generation, which has traditionally relied on ControlNet-like (Zhang et al., 2023) methods with input-specific encoders and adapters, RINO establishes a new zero-shot state of the art under existing evaluation protocols. Visually, RINO's understanding and generation results also receive high scores in human preference studies, suggesting that RGB-based unification has strong potential for human interaction and may serve as a transferable "language" for vision.

Overall, this work provides strong evidence for the feasibility of RGB as a unified interface for vision. We present an initial attempt toward building a large-scale, multi-task unified vision-language system, and observe highly promising signs of broad task transfer under a single RGB-based formulation. At this stage, our RINO models still have limitations: their strong transfer ability relies on generic image editing backbones like Qwen-Image-Edit (Wu et al., 2025) and FireRed (FireRed Team, 2025), whose web-scale pretraining data may already contain diverse visual signals such as segmentation masks and depth maps. However, such signals likely occupy only a very small portion of the pretraining distribution, which may lead to a performance gap between RINO and in-domain expert models on some specific applications. We refer to the current fully zero-shot version as **RINO-Zero**. In future work, we will instruction-tune existing image editing models with standard image editing data mixed with a proportion of RGB-formatted non-natural visual data, such as segmentation masks and depth maps, so that the RINO framework is introduced during training across a wider range of tasks. We hope this work can provide useful insights for future unified vision-language systems and help extend this paradigm to broader vision domains, including 3D vision, video, and world models.

## 2 Method: Unified RGB Interface for Vision

We formulate general vision-language understanding and generation tasks under a unified RGB/text-to-RGB/text paradigm. Given an input image $x_{\text{img}}$ and an input text prompt $x_{\text{text}}$, a multimodal image editing model $F$ produces an output image $y_{\text{img}}$ and, optionally, an output text response $y_{\text{text}}$:

$$y_{\text{img}}, y_{\text{text}} = \mathcal{F}(x_{\text{img}}, x_{\text{text}}), \tag{1}$$

where $\mathcal{F}$ is a generic pretrained image editing model such as Qwen-Image-Edit (Wu et al., 2025). In our framework, we do not introduce task-specific heads, encoders, decoders, or adapter parameters. Instead, all visual information is represented in RGB format and processed by the same image encoder and decoder inside $\mathcal{F}$. This allows natural images, segmentation masks, depth maps, pose maps, edge maps, and other structured visual signals to share the same visual interface. We illustrate our overall framework in Figure 1.

**For visual understanding tasks**, the input $x_{\text{img}}$ is a natural image, and the input $x_{\text{text}}$ specifies the task to be performed. The output image $y_{\text{img}}$ is an RGB representation of the target structured visual information. For example, in semantic segmentation, $y_{\text{img}}$ is an RGB segmentation mask; in depth estimation, $y_{\text{img}}$ is an RGB-formatted depth map. The optional text output $y_{\text{text}}$ can be used to explain the prediction or describe the generated visual result. When evaluating $y_{\text{img}}$, we use a parameter-free converter to transform the RGB output into the corresponding task-specific format. For example, an RGB segmentation map can be converted into class labels through a predefined color mapping, and an RGB depth visualization can be converted into a depth map through a predefined decoding rule. These converters are only used before or after model inference. All visual reasoning is performed inside the base model $\mathcal{F}$ through the unified RGB interface.

**For visual generation tasks**, the input $x_{\text{img}}$ can be an RGB-formatted visual condition, such as a semantic layout, a pose skeleton, a depth map, an edge map, or a mask. The text prompt $x_{\text{text}}$ describes the desired generation or editing behavior. The model then generates an output image $y_{\text{img}}$, which is usually a natural image that follows the given visual condition and text instruction. This formulation treats conditioned generation as the dual problem of visual understanding. In understanding, the model maps a natural image to an RGB-formatted structured prediction. In generation, the model maps an RGB-formatted structured condition back to a natural image. For example, if semantic segmentation can be formulated as image-to-RGB-mask prediction, then segmentation-conditioned image generation can be formulated as RGB-mask-to-image generation. Similarly, pose estimation can be viewed as image-to-RGB-pose prediction, while pose-guided generation can be viewed as RGB-pose-to-image generation.

**Base models.** We run RINO-Zero on three open-sourced image editors: Qwen-Image-Edit (Wu et al., 2025), LongCat-Image-Edit (Ma et al., 2025) and FireRed-Image-Edit (FireRed Team, 2025). Qwen-Image-Edit builds on Qwen-Image, a 20B-parameter multimodal diffusion transformer (MMDiT) that couples a Qwen2.5-VL vision-language model with a VAE for image tokens. LongCat-Image-Edit is a compact 6B bilingual (Chinese-English) editor from Meituan, built on a hybrid MMDiT / single-stream DiT design with a Qwen2.5-VL text encoder and tuned for efficiency at a smaller size. FireRed-Image-Edit, from the FireRed team, extends an open-source text-to-image foundation model with a double-stream multimodal-diffusion design for high-fidelity, instruction-following edits. For all three we use the released weights as a black-box RGB-to-RGB editor, without adding, removing, or fine-tuning any layers.

## 3 Experiments

### 3.1 Understanding Tasks: Unifying Outputs as RGB

**Depth Estimation.** For depth estimation, we prompt the image editor to repaint the input image as a grayscale depth visualization, where nearby surfaces are brighter and distant surfaces are darker. We then recover a dense relative depth map from the per-pixel luminance of the generated image. In this task, we evaluate with Qwen-Image-Edit (Wu et al., 2025) and LongCat-Image-Edit (Ma et al., 2025) in a zero-shot setting with a fixed sampling configuration. Since a generative editor produces relative depth with unknown scale and offset, we follow the standard affine-invariant evaluation protocol. For each image, we fit a scale and shift to the ground-truth depth by least squares before computing $\delta_1$ and AbsRel. Table 1 follows the *disparity*-space, or

**Table 1** Zero-shot monocular depth estimation under a disparity-space affine-invariant protocol. Specialist scores are quoted from Yang et al. (2024b) (ViT-L); $\delta_1 \uparrow$/AbsRel$\downarrow$. **Bold**: best per metric.

| | NYUv2 | | KITTI | | DIODE-indoor | |
|---|---|---|---|---|---|---|
| Method | $\delta_1$ | AbsRel | $\delta_1$ | AbsRel | $\delta_1$ | AbsRel |
| *Relative-depth specialists* | | | | | | |
| MiDaS V3.1 (Birkl et al., 2023) | 0.980 | 0.048 | 0.850 | 0.127 | 0.942 | 0.075 |
| Depth Anything V1 (Yang et al., 2024a) | **0.981** | **0.043** | **0.947** | 0.076 | **0.952** | **0.066** |
| Depth Anything V2 (Yang et al., 2024b) | 0.979 | 0.045 | 0.946 | **0.074** | **0.952** | **0.066** |
| *Generative image editors* | | | | | | |
| Qwen-Image-Edit | 0.927 | 0.086 | 0.901 | 0.093 | 0.938 | 0.076 |
| LongCat-Image-Edit | 0.840 | 0.130 | 0.849 | 0.130 | 0.849 | 0.129 |

**Table 2** The same editors under the *linear* depth-space affine-invariant protocol of Liu et al. (2026) (per-image least-squares scale + shift on GT depth). †Vision Banana (Gabeur et al., 2026) reports *raw metric* depth with no alignment, shown only as a reference and not directly comparable. $\delta_1 \uparrow$/AbsRel$\downarrow$. **Bold**: best editor per metric.

| | NYUv2 | | KITTI | | DIODE-indoor | | iBims-1 | |
|---|---|---|---|---|---|---|---|---|
| Method | $\delta_1$ | AbsRel | $\delta_1$ | AbsRel | $\delta_1$ | AbsRel | $\delta_1$ | AbsRel |
| *Metric reference (instruction-tuned)* | | | | | | | | |
| Vision Banana† | 0.948 | 0.081 | 0.915 | 0.107 | 0.917 | 0.108 | 0.934 | 0.078 |
| *Generative image editors (linear affine-invariant)* | | | | | | | | |
| Qwen-Image-Edit | 0.792 | 0.156 | 0.413 | 0.329 | **0.872** | **0.122** | 0.824 | 0.133 |
| LongCat-Image-Edit | **0.832** | **0.135** | **0.454** | **0.325** | 0.825 | 0.146 | **0.843** | **0.129** |

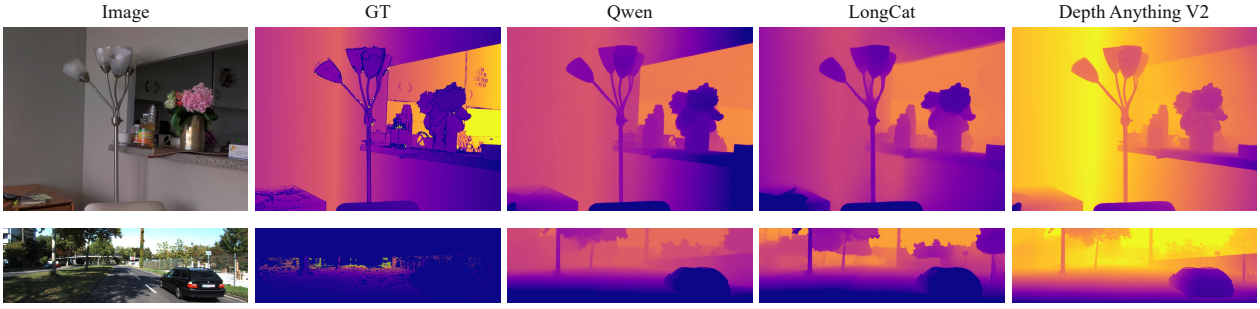


**Figure 2** Qualitative zero-shot depth estimation on DIODE-indoor (top) and KITTI (bottom).

inverse-depth, convention used by MiDaS and Depth Anything (Ranftl et al., 2022; Yang et al., 2024a), while Table 2 follows the *linear* depth-space convention used by Marigold (Ke et al., 2024).

Table 1 compares the two editors with three relative-depth specialists in disparity space: MiDaS V3.1 (Birkl et al., 2023), Depth Anything V1 (Yang et al., 2024a), and Depth Anything V2 (Yang et al., 2024b), with specialist results quoted from Yang et al. (2024b). Although neither editor is trained for depth estimation or equipped with any depth-specific parameters, Qwen-Image-Edit performs surprisingly close to dedicated models. It remains within a few $\delta_1$ points of the specialists on DIODE-indoor (Vasiljevic et al., 2019) (0.938 vs. 0.952), and shows a roughly five-point gap on NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2012). Table 2 reports the same comparison in linear depth space, where performance drops on wide-range scenes such as KITTI and iBims-1 (Koch et al., 2018). We include Vision Banana (Gabeur et al., 2026) as a reference in this table, since it predicts raw metric depth without per-image alignment and is therefore not directly comparable to the affine-invariant editor outputs. Qualitative visualizations are shown in Figure 2. We observe that RINO's prediction can clearly reflect the depth layout of input images.

**Table 3** Zero-shot surface-normal estimation on NYUv2, iBims-1, and DIODE-indoor. We report Mean↓ and median↓ angular error (degrees). **Bold**: best editor per metric.

| | NYUv2 | | iBims-1 | | DIODE-indoor | |
|---|---|---|---|---|---|---|
| Method | Mean | Median | Mean | Median | Mean | Median |
| *Surface-normal specialists (quoted; see caption)* | | | | | | |
| DSINE (Bae and Davison, 2024) | 16.4 | 8.4 | 17.1 | 6.1 | 18.45 | 13.87 |
| Marigold (Ke et al., 2024) | 20.86 | 11.13 | 18.46 | 8.44 | 16.67 | 12.08 |
| StableNormal (Ye et al., 2024) | 19.71 | 10.53 | 17.25 | 8.06 | 13.70 | 9.46 |
| Lotus-2 (He et al., 2025) | 16.9 | — | 15.4 | — | 18.58 | — |
| Vision Banana (Gabeur et al., 2026) | 17.78 | 8.88 | — | — | 13.82 | 11.56 |
| *Generative image editors (ours, zero-shot, RGB→RGB)* | | | | | | |
| Qwen-Image-Edit | 20.21 | 13.46 | 20.66 | 12.03 | 21.99 | 18.14 |
| FireRed-Image-Edit | **17.73** | **10.99** | **18.62** | **10.38** | **17.25** | **13.64** |

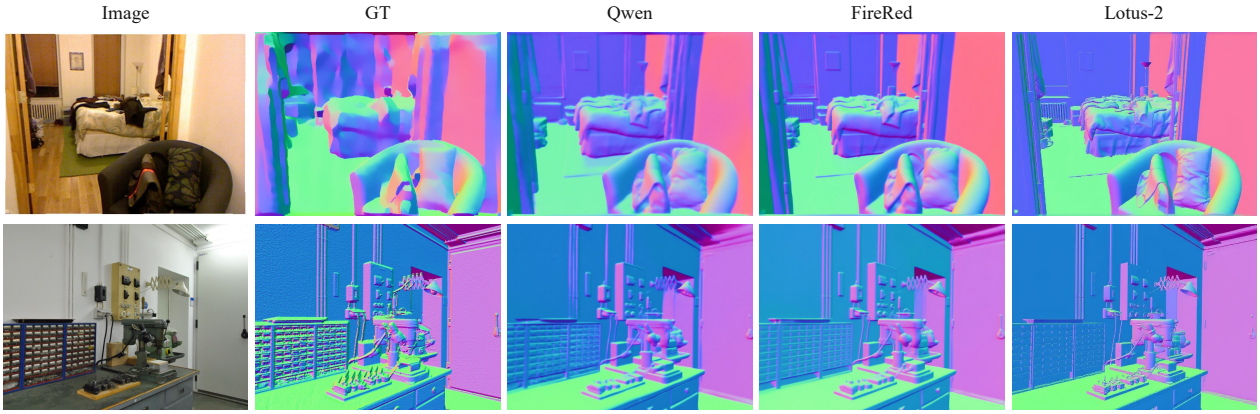


**Figure 3** Qualitative zero-shot surface normal estimation on NYUv2 (top) and iBims-1 (bottom).

**Surface Normal Estimation.** Surface normal estimation follows the same RGB-to-RGB paradigm as depth estimation. We prompt the editors to repaint the input scene as a standard normal map, where surface orientation is encoded by color, and decode the generated RGB image back into per-pixel unit normals. We evaluate Qwen-Image-Edit and FireRed-Image-Edit in a zero-shot setting. The image editors output normals in an unknown coordinate convention, which we resolve once following Liu et al. (2026), and report the standard mean and median angular error (degrees). DSINE (Bae and Davison, 2024) is scored through the same pipeline as a specialist reference. Table 3 reports surface normal estimation results on NYUv2, iBims-1, and DIODE-indoor. FireRed, a generic image editor without normal-specific training or parameters, is the strongest editor across all three datasets. It only slightly trails the DSINE specialist in mean angular error, although DSINE remains sharper at the pixel level, as reflected by its lower median error. These results suggest that a zero-shot image editor can approach a dedicated normal estimation model in average surface orientation prediction, while still leaving room for improvement in local precision. Qualitative visualizations are shown in Figure 3.

**Semantic Segmentation.** We follow Liu et al. (2026), where only the categories present in the ground-truth of the current image are included in the prompt. To mitigate common generation artifacts such as color jitter and noisy pixels, we apply a lightweight parameter-free refinement stage after color decoding, consisting of local majority filtering and small connected-component removal. To handle the large label space of ADE20K, we introduce a hierarchical segmentation strategy: the model first predicts masks for 10 super-class, and fine-grained categories are subsequently generated within their corresponding regions. Unless otherwise specified, all image editing backbones use 12 denoising steps, CFG = 5.0, and the same refinement pipeline. LongCat-Image-Edit employs 50 denoising steps due to its lower generation stability. We use three standard

**Table 4** Semantic segmentation results on Cityscapes and Pascal VOC. **Bold** and underline indicate the best and second-best results among generative image editors.

| Method | Pascal VOC | | | Cityscapes | | |
|---|---|---|---|---|---|---|
| | mIoU ↑ | mAcc ↑ | aAcc ↑ | mIoU ↑ | mAcc ↑ | aAcc ↑ |
| *Zero-shot Segmentation* | | | | | | |
| FC-CLIP (Yu et al., 2023) | — | — | — | 56.29 | 65.42 | 78.46 |
| MaskCLIP (Dong et al., 2023; Barsellotti et al., 2024) | 38.85 | 45.21 | 71.63 | — | — | — |
| GroupViT (Xu et al., 2022a) | 52.37 | 56.97 | 79.28 | — | — | — |
| OVSegmentor (Xu et al., 2023) | 53.82 | 58.36 | 80.51 | — | — | — |
| *Generative image editors (zero-shot, RGB→RGB)* | | | | | | |
| Vision Banana (Gabeur et al., 2026) | — | — | — | 69.90 | — | — |
| Direct Editing Baseline (Liu et al., 2026) | — | — | — | 23.02 | 42.63 | 64.14 |
| Ours (Qwen-Image-Edit) | 44.69 | 48.00 | **85.17** | 24.30 | **44.38** | **66.59** |
| Ours (FireRed-Image-Edit) | 45.17 | 61.47 | 80.82 | 21.11 | 42.70 | 58.00 |
| Ours (LongCat-Image-Edit) | **49.68** | **66.90** | 83.53 | 13.42 | 26.72 | 49.23 |

**Table 5** Semantic segmentation results on ADE20K under both the 10-class coarse setting and the standard 150-category setting. **Bold** and underline indicate the best and second-best results among generative image editors.

| Method | ADE20K (10 classes) | | | ADE20K (150 categories) | | |
|---|---|---|---|---|---|---|
| | mIoU ↑ | mAcc ↑ | aAcc ↑ | mIoU ↑ | mAcc ↑ | aAcc ↑ |
| *Zero-shot Segmentation* | | | | | | |
| MaskCLIP (Dong et al., 2023) | 39.26 | 50.52 | 65.66 | 23.78 | 31.54 | 50.27 |
| FC-CLIP (Yu et al., 2023) | 42.17 | 53.11 | 67.05 | 34.13 | 40.86 | 59.62 |
| *Generative image editors (ours, zero-shot, RGB→RGB)* | | | | | | |
| Qwen-Image-Edit | **40.37** | **53.51** | **67.35** | 12.57 | 20.93 | **44.85** |
| FireRed-Image-Edit | 34.76 | 46.79 | 54.77 | **13.61** | **22.80** | 38.95 |
| LongCat-Image-Edit | 19.33 | 30.02 | 26.37 | 7.43 | 14.60 | 26.79 |

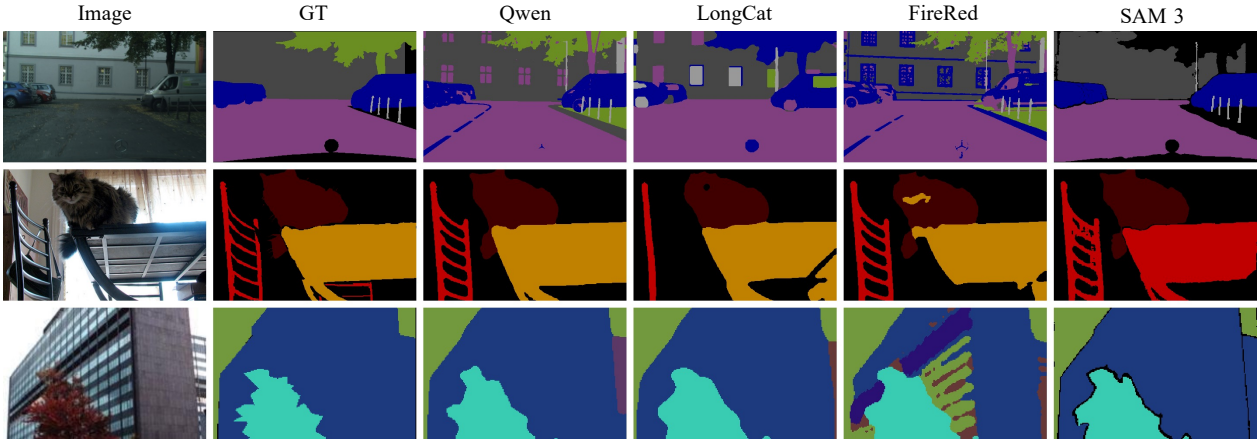


**Figure 4** Zero-shot semantic segmentation on Cityscapes (top), Pascal VOC (middle), and ADE20K (bottom).

segmentation metrics: mean Intersection over Union (mIoU), mean class accuracy (mAcc), and all-pixel accuracy (aAcc). Quantatitive results are reported on the validation sets of Cityscapes (Cordts et al., 2016), Pascal VOC (Everingham et al., 2010), and ADE20K (Zhou et al., 2017).

Table 4 shows that RINO achieves semantic segmentation performance comparable to zero-shot specialist

**Table 6** Object detection and instance segmentation on COCO (box and mask AP, %). Editors are run by us zero-shot under an oracle-class, oracle-count per-class silhouette protocol, the same blobs giving boxes and masks (all-points AP, macro-averaged over 80 classes). **Bold**: best editor.

| | box AP | | mask AP | |
|---|---|---|---|---|
| Method | $AP_{50}$ | AP | $AP_{50}$ | AP |
| *Supervised detect + segment models (AP quoted from their papers)* | | | | |
| Mask DINO (Li et al., 2023a) | — | 50.5 | — | 46.0 |
| Cascade Mask R-CNN (Cai and Vasconcelos, 2021) | 61.7 | 43.3 | 58.6 | 37.1 |
| Mask R-CNN (He et al., 2017) | 63.5 | 42.8 | 60.5 | 38.3 |
| Mask2Former (Cheng et al., 2022) | 69.4 | 49.3 | 69.4 | 46.1 |
| *Generative image editors (ours, zero-shot, RGB→RGB)* | | | | |
| Qwen-Image-Edit | **16.3** | **9.4** | **13.9** | **7.2** |
| FireRed-Image-Edit | 14.9 | 8.9 | 12.9 | 6.9 |
| LongCat-Image-Edit | 6.3 | 3.2 | 4.9 | 2.4 |

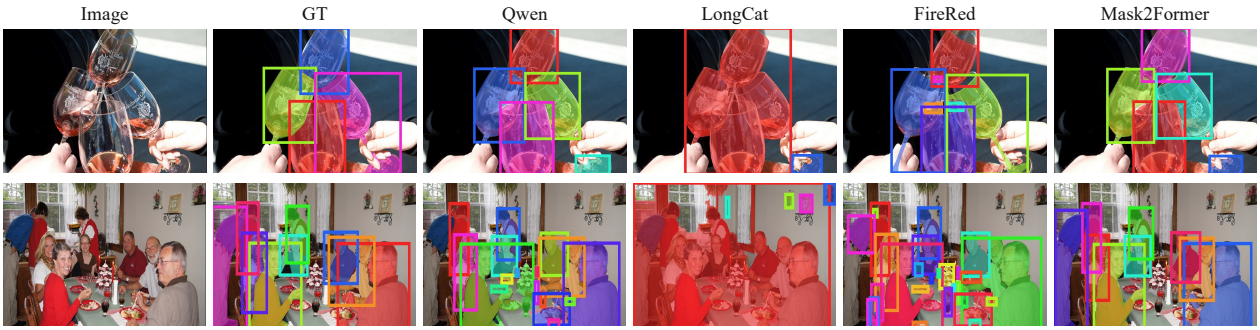


**Figure 5** Qualitative object detection and instance segmentation on COCO val2017.

models, with mAcc and aAcc even surpassing segmentation experts on PASCAL VOC. Table 5 further evaluates segmentation at both coarse and fine granularities on ADE20K. We observe that under the coarse setting with 10 classes, RINO can even outperform the semantic segmentation specialist MaskCLIP. These results suggest that RINO has strong potential for dense prediction: the model already captures rich image semantics, but may require more domain-specific knowledge to better align its predictions with fine-grained category names, such as the 150 classes in ADE20K. Visualization results are shown in Figure 4.

**Object Detection and Instance Segmentation.** Object detection and instance segmentation are challenging vision tasks as they require not only recognizing per-pixel semantics but also distinguishing different instances within the same category. As a result, open-vocabulary zero-shot models often struggle on datasets such as COCO, which contains 80 object categories. In RINO, we perform both tasks through per-class silhouette painting. For each present class, we prompt the editor to paint its instances as solid white blobs on a black background, while also specifying the expected number of instances. We then threshold the output and split touching blobs using morphological opening. Each connected component is treated as one instance: the blob defines its mask, and its tight bounding box defines the detection. Thus, a single prediction can be evaluated for both object detection and instance segmentation. The queried classes and instance counts are taken from the ground truth, i.e., an oracle-class and oracle-count setting, to isolate localization ability from open-set recognition. We evaluate Qwen-Image-Edit (Wu et al., 2025), FireRed-Image-Edit (FireRed Team, 2025), and LongCat-Image-Edit (Ma et al., 2025) in a zero-shot manner on COCO val2017.

As shown in Table 6 and Figure 5, although these tasks are highly challenging for zero-shot models, RINO still produces visually meaningful predictions, achieving 16.3% box $AP_{50}$ and 13.9% mask $AP_{50}$. These results remain below in-domain specialist models, such as Mask R-CNN with 63.5% box $AP_{50}$, but they demonstrate that a zero-shot image editing model can already reach a non-trivial level of localization and instance-level

**Table 7** Zero-shot panoptic segmentation on the Cityscapes, ADE20K and COCO validation sets. Supervised specialists report Panoptic Quality (PQ, %) as in their papers; editor cells report Segmentation Quality (SQ, %) and additionally give (PQ/RQ) in parentheses (our measurements). **Bold**: best SQ results.

| Method | Cityscapes | ADE20K | COCO |
|---|---|---|---|
| *Supervised specialists* | | | |
| EoMT (Kerssies et al., 2025) | — | 51.7 | 58.3 |
| ViT-P (Shahabodini et al., 2025) | 70.8 | 51.9 | 58.0 |
| OneFormer (Jain et al., 2023) | 67.2 | 49.8 | 57.9 |
| Mask2Former (Cheng et al., 2022) | 66.6 | 48.1 | 57.8 |
| *Generative image editors* | | | |
| Qwen-Image-Edit | **64.8** (13.6/17.4) | **71.5** (12.3/15.9) | 74.3 (11.9/15.5) |
| FireRed-Image-Edit | 55.9 ( 8.7/11.2) | 67.2 ( 9.4/12.5) | 70.7 ( 9.1/12.1) |
| LongCat-Image-Edit | 55.4 ( 1.0/ 1.4) | 67.6 ( 6.7/ 8.4) | **75.2** ( 6.9/ 8.8) |

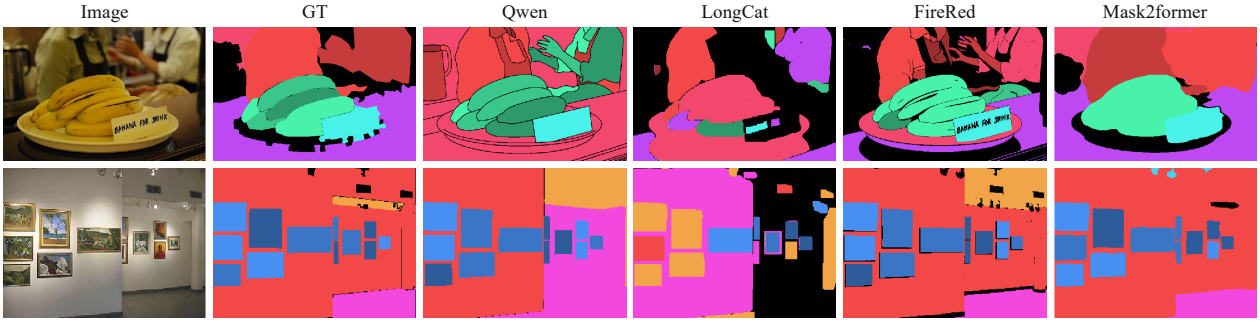


**Figure 6** Qualitative zero-shot panoptic segmentation on COCO (top) and ADE20K (bottom).

prediction on COCO. Figure 5 further suggests that the relatively low quantitative scores are often caused by task-specific failure modes such as over-detection or over-segmentation, as observed in the Qwen and FireRed examples. We believe these issues can be substantially mitigated with a small amount of task-specific fine-tuning, and expect future versions of RINO to match the performance of specialist models.

**Panoptic Segmentation.** We prompt the editor to paint each region with a flat color, using one named color per class and black for the background. We then decode the output by nearest-color matching: stuff classes are treated as single segments, while thing classes are split into connected components. Following the zero-shot oracle-class protocol, only the classes present in each image are specified in the prompt. We evaluate Qwen-Image-Edit (Wu et al., 2025), FireRed-Image-Edit (FireRed Team, 2025), and LongCat-Image-Edit (Ma et al., 2025). We report Panoptic Quality (PQ) and its two components, Segmentation Quality (SQ) and Recognition Quality (RQ), where $\mathrm{PQ} = \mathrm{SQ} \times \mathrm{RQ}$. PQ is the standard metric for panoptic segmentation: SQ measures the mask quality of matched regions, while RQ measures recognition and matching quality.

Table 7 reports results on Cityscapes (Cordts et al., 2016), ADE20K (Zhou et al., 2017), and COCO (Lin et al., 2014). Among the editors, Qwen-Image-Edit performs best, followed by FireRed and LongCat. Notably, despite no panoptic-specific training, the editors already show strong zero-shot spatial segmentation ability. Their predicted masks are often well aligned with object and region boundaries, leading to high SQ scores (Qwen SQ 65–74). The main limitation instead lies in recognition and instance association, as reflected by lower RQ scores (Qwen RQ 15–17). In many cases, coherent regions are correctly separated but assigned to wrong semantic categories, or split into multiple segments due to task-specific decoding ambiguities. This issue is most pronounced for *thing* classes, where adjacent instances and small objects are often merged or missed; on Cityscapes, Qwen reaches 22.4 PQ on *stuff* but only 1.5 PQ on *things*. These results suggest that RINO can already produce structurally meaningful panoptic predictions zero-shot, while the remaining gap is largely driven by task-specific recognition and instance-matching factors rather than spatial segmentation quality. This is also supported by the visualizations in Figure 6.

**Table 8** Zero-shot 2D human pose estimation on COCO val2017 single-person crops. PCK is bbox-normalized and reported both in absolute image coordinates (*raw*) and after per-image similarity alignment (*PA*), which isolates pose structure by removing scale and translation. Dedicated pose estimators are evaluated under the same top-down protocol for direct comparison.

| Method | raw PCK@0.05 | PA PCK@0.05 | PA PCK@0.2 |
|---|---|---|---|
| *2D-pose specialists (run under our protocol)* | | | |
| ViTPose (Xu et al., 2022b) | 0.868 | 0.889 | 0.984 |
| Keypoint R-CNN (He et al., 2017) | 0.732 | 0.765 | 0.919 |
| YOLO-pose (Maji et al., 2022) | 0.730 | 0.782 | 0.953 |
| *Mesh-recovery specialist* | | | |
| SAM 3D Body (Yang et al., 2026b) | 0.868 | — | — |
| *Generative image editors* | | | |
| LongCat-Image-Edit | 0.233 | 0.282 | 0.767 |
| Qwen-Image-Edit | 0.060 | 0.235 | 0.774 |
| FireRed-Image-Edit | 0.047 | 0.243 | 0.795 |

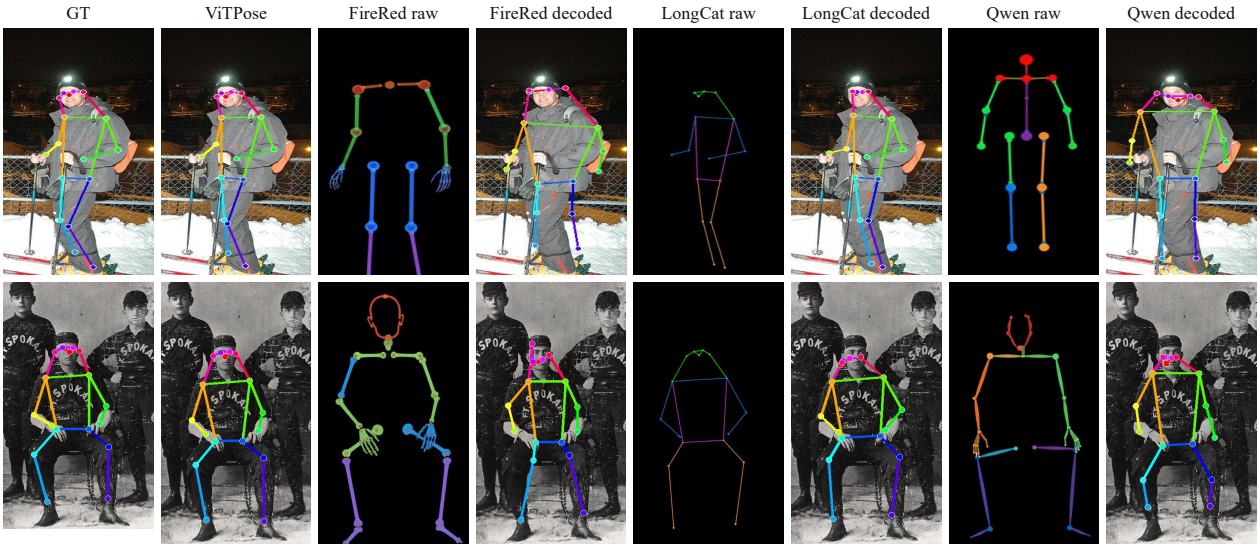


**Figure 7** Qualitative zero-shot 2D human pose estimation.

**Human Pose Estimation.** Following the top-down protocol, each person is cropped using its ground-truth box and letterboxed to a fixed canvas. We prompt the editor to repaint the crop as an OpenPose-style (Cao et al., 2021) colored skeleton on a black background, and then decode COCO-17 keypoints from the output. Since image editors do not reliably reproduce a fixed color palette, we use a color-agnostic decoder: joint candidates are extracted as peaks of the foreground distance transform and labeled by matching to a canonical upright skeleton template. No keypoint head or auxiliary parameters are introduced. We evaluate Qwen-Image-Edit (Wu et al., 2025), FireRed-Image-Edit (FireRed Team, 2025), and LongCat-Image-Edit (Ma et al., 2025) in a zero-shot manner on COCO val2017 single-person crops. Since a generative editor may predict pose up to an unknown similarity transform, we report bbox-normalized PCK@$\tau$ under two protocols: *raw*, in absolute image coordinates, following dedicated estimators, and *PA*, after a per-image scale-and-translation fit to the ground truth, which isolates pose structure.

Table 8 reports results on COCO val2017. The editors already produce non-trivial zero-shot pose predictions, often recovering coherent human-body layouts and plausible limb structures. Although precise keypoint localization remains challenging, the generated skeletons are visually meaningful and can be decoded into COCO-17 keypoints without any pose-specific head or training. The improved performance under similarity

**Table 9** Referring expression comprehension (REC) and segmentation (RES) on RefCOCOg (UMD split) val. REC: box Prec@0.5; RES: cIoU/mIoU. **Bold**: best editor per metric.

| | REC | RES | |
|---|---|---|---|
| Method | Prec@0.5 | cIoU | mIoU |
| *Supervised grounding specialists* | | | |
| LAVT (Yang et al., 2022) | — | 61.2 | — |
| UNINEXT (Yan et al., 2023) | 88.7 | 74.7 | — |
| OneRef (Xiao et al., 2024) | 88.1 | — | 73.2 |
| *Generative image models* | | | |
| Vision Banana (Gabeur et al., 2026) | — | 73.8 | — |
| Qwen-Image-Edit | **51.8** | **40.3** | **45.3** |
| LongCat-Image-Edit | 47.7 | 35.0 | 38.7 |
| FireRed-Image-Edit | 47.6 | 35.7 | 41.6 |

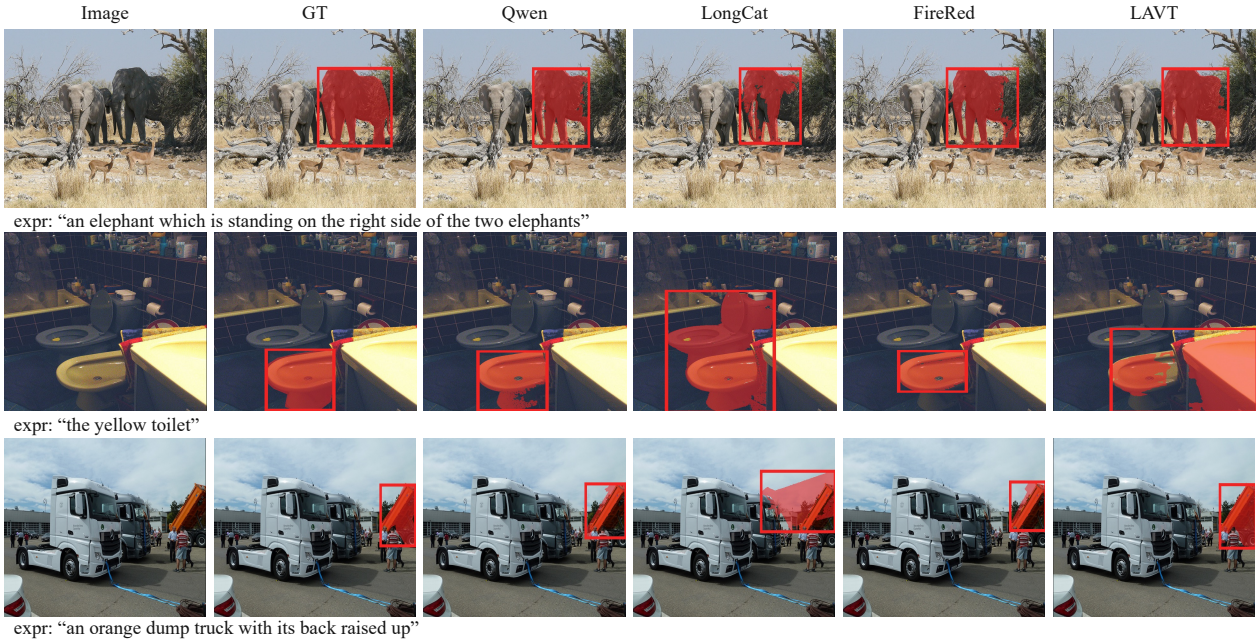


**Figure 8** Qualitative referring expression comprehension and segmentation on RefCOCOg.

alignment further suggests that these models capture the coarse structure of human pose, even when absolute localization is imperfect. These results indicate that RGB skeleton prediction is a viable interface for pose estimation under the RINO framework. With lightweight task-specific fine-tuning, we expect this ability to become substantially stronger and more reliable. Qualitative visualizations are shown in Figure 7.

**Referring Expression Comprehension and Segmentation.** We perform language grounding through image editing. For each referring expression, we prompt the editor to cover the referred object in solid red while leaving all other pixels unchanged. We then compute the difference between the edited output and the input image, and keep the largest changed connected component as the prediction: its pixels define the segmentation mask, and its tight bounding box defines the comprehension result. The editor only receives the image and the expression, without access to boxes, classes, or instance counts, and no grounding head is introduced. We evaluate Qwen-Image-Edit, FireRed-Image-Edit, and LongCat-Image-Edit zero-shot on the RefCOCOg UMD validation split (Mao et al., 2016; Nagaraja et al., 2016). We report the standard metrics for both subtasks: box Prec@0.5 for referring expression comprehension (REC), where a prediction is correct if its box IoU is at least 0.5, and cumulative IoU (cIoU) for referring expression segmentation (RES), computed as dataset-pooled

**Table 10** Open-vocabulary instance segmentation on the *attributes* subset of SA-Co/Gold (cgF1 = 100·pmF1·IL_MCC) Note that pmF1 is gate-independent and we obtain comparable results as specialist models. SAM 3 is supervised; Vision Banana adds an MLLM presence gate. **Bold:** best editor results.

| Method | pmF1 | IL_MCC | cgF1 |
|---|---|---|---|
| *Supervised specialist (quoted, not zero-shot transfer)* | | | |
| SAM 3 (Carion et al., 2025) | 72.0 | 0.76 | 54.9 |
| *Zero-shot transfer* | | | |
| gDino-T (Liu et al., 2024) | 47.3 | 0.29 | 13.8 |
| Vision Banana + Gemini (Gabeur et al., 2026) | 68.7 | 0.86 | 58.8 |
| *Generative image editors (ours, zero-shot, RGB→RGB)* | | | |
| Qwen-Image-Edit | **39.8** | 0.04 | **1.7** |
| FireRed-Image-Edit | 30.9 | 0.01 | 0.2 |
| LongCat-Image-Edit | 17.3 | **0.09** | 1.6 |

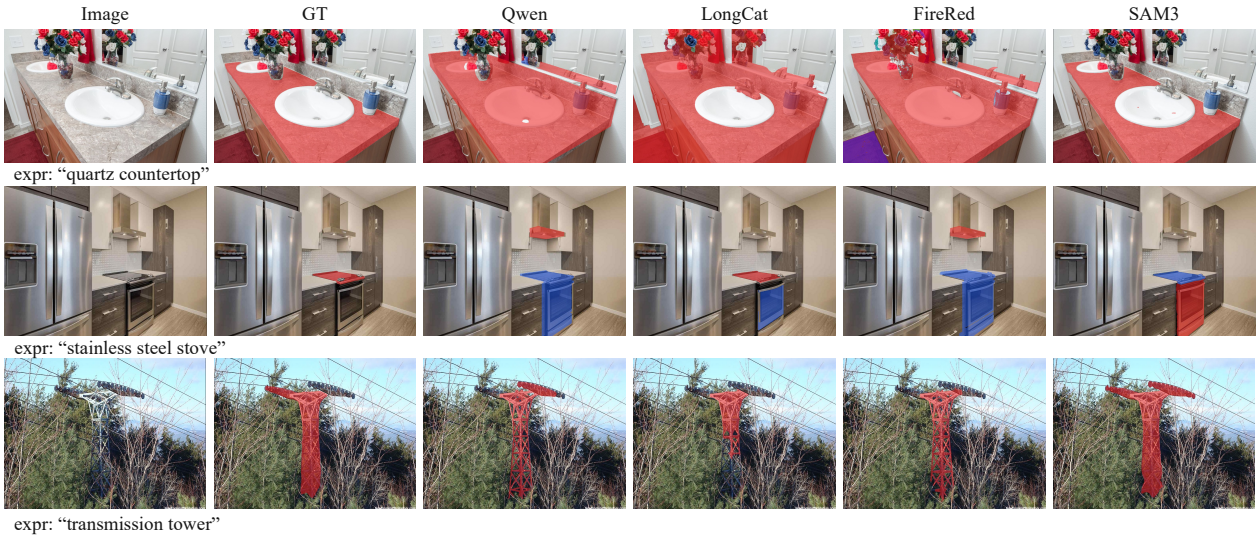


**Figure 9** Qualitative open-vocabulary instance segmentation on SA-Co/Gold.

intersection-over-union. Missed predictions are assigned zero score.

As shown in Table 9, the open-source editors achieve meaningful zero-shot grounding from free-form language to image regions. Although their segmentation accuracy is still below Vision Banana (Gabeur et al., 2026), whose cIoU is comparable to supervised grounders such as UNINEXT and OneRef, they can already localize and segment referred objects without any grounding-specific head, box supervision, or task-specific fine-tuning. This is a substantially harder setting than oracle-class segmentation, since the model must jointly parse the referring expression, identify the target instance, and express the grounded region as an RGB edit. The results suggest that generic image editors already possess a non-trivial ability to connect language with pixel-level visual grounding through the RGB editing interface. Qualitative examples in Figure 8 further show that many predictions are visually interpretable and spatially well formed, indicating a promising direction for improving referring understanding under the RINO framework.

**Open-Vocabulary Instance Segmentation (SA-Co).** For each query, we prompt the editor once to repaint the image as a binary mask: all instances matching the phrase are painted pure black, all other regions are painted pure white, and the output should be fully white if the phrase is absent. We take black pixels as foreground and split them into instances using connected components, dropping small components and scoring each instance by area. The model is given no class list or instance count; the query is a single free-form noun phrase that may or may not appear in the image, requiring the editor to perform both concept localization

and presence detection. No task-specific head is introduced. We evaluate Qwen-Image-Edit (Wu et al., 2025), FireRed-Image-Edit (FireRed Team, 2025), and LongCat-Image-Edit (Ma et al., 2025) zero-shot on the attributes split of SA-Co/Gold (Carion et al., 2025). We report the three official SA-Co metrics: pmF1, the positive-only micro mask-F1 averaged over IoU thresholds from 0.5 to 0.95; IL_MCC, the Matthews correlation coefficient for per-query present/absent prediction; and $\text{cgF1} = 100 \cdot \text{pmF1} \cdot \text{IL}_{\text{M}}\text{CC}$, the headline score combining segmentation quality and presence detection. We report pmF1 and cgF1 as percentages.

Table 10 reports results on the attributes subset. Qwen-Image-Edit performs best among the editors, followed by FireRed and LongCat. While the editors do not yet match dedicated open-vocabulary segmentation models in pmF1, they already produce coherent zero-shot masks for free-form visual concepts without any class list, detector, or segmentation head. The main weakness is open-set presence detection: since the editors tend to paint a mask even when the queried phrase is absent, IL_MCC remains close to zero, which in turn suppresses the headline cgF1. This limitation reflects the absence of an explicit presence classifier rather than a complete failure of RGB-space segmentation; indeed, even Vision Banana relies on an external MLLM gate to obtain a meaningful IL_MCC. These results suggest that generic image editors can localize named concepts through the RGB editing interface, but reliable absent-case rejection remains an important direction for future improvement. Qualitative visualizations are shown in Figure 9.

## 3.2 Generation Tasks: Unifying Inputs as RGB

**Table 11** Depth-conditioned image generation on the MultiGen-20M depth validation split. External rows are paper-reported anchors from ControlNet++ and ControlAR; our editor rows are zero-shot RGB-in/RGB-out editing runs under a shared prompt, seed 0, `steps=8`, and `CFG=2.5`. RMSE-255↓ measures DPT-Large depth consistency, FID↓ measures image quality, and CLIPScore↑ measures text alignment.

| Method | N | RMSE-255↓ | FID↓ | CLIPScore↑ |
|---|---|---|---|---|
| *Trained controllable generation models* | | | | |
| ControlNet++ (Li et al., 2024a) | 5000 | 28.32 | 16.66 | 32.09 |
| ControlAR (Li et al., 2024b) | 5000 | 29.01 | 14.61 | – |
| ControlNet (SD1.5) (Zhang et al., 2023) | 5000 | 35.90 | 17.76 | 32.45 |
| GLIGEN (Li et al., 2023b) | 5000 | 38.83 | 18.36 | 31.75 |
| T2I-Adapter (Mou et al., 2023) | 5000 | 48.40 | 22.52 | 31.46 |
| Uni-ControlNet (Zhao et al., 2023) | 5000 | 40.65 | 20.27 | 31.66 |
| UniControl (Qin et al., 2023) | 5000 | 39.18 | 18.66 | 32.45 |
| *Generative image editors (ours, zero-shot, RGB depth input)* | | | | |
| Qwen-Image-Edit | 5000 | **33.83** | **19.44** | 30.15 |
| FireRed-Image-Edit | 5000 | 37.55 | 21.43 | **30.34** |
| LongCat-Image-Edit | 5000 | 36.50 | 27.18 | 28.09 |

**Depth-Conditioned Image Generation.** We evaluate depth-to-image generation as a direct test of RGB-in/RGB-out controllable generation. Following the MultiGen-20M depth protocol, each validation sample provides an RGB depth map and a text caption. RINO passes the depth map to an image editor as an ordinary image condition and instructs it to synthesize the corresponding natural image. No depth encoder, ControlNet branch, adapter, or task-specific decoder is added; the model only sees an RGB condition image and a natural-language instruction. For the main paper-facing comparison, we evaluate Qwen-Image-Edit, FireRed-Image-Edit, and LongCat-Image-Edit on the same 5,000-image validation split with the same prompt, seed, and inference setting: `steps=8`, `CFG=2.5`, and a shared short quality prompt. We follow the ControlNet++ evaluation protocol for depth-conditioned generation (Li et al., 2024a). Condition fidelity is measured by extracting depth from generated RGB images with DPT-Large (Ranftl et al., 2021) and computing RMSE in 0–255 depth space against the input condition. Image quality is measured by FID against the real validation images, and text alignment by CLIPScore. Lower RMSE and FID are better; higher CLIPScore is better. The external rows are paper-reported anchors rather than locally re-scored outputs. We also note a resolution caveat: the editor outputs in our pipeline are generated at their native 1024-pixel resolution from 512-pixel MultiGen control maps, whereas the public depth-control baselines are reported at 512 pixels. A downsample-to-512 sanity check preserves the same qualitative conclusion, but the most directly aligned cross-method metric is


**Figure 10** Qualitative depth-conditioned generation on MultiGen-20M. The same RGB depth maps and text prompts are given to all three zero-shot editors. The three examples show that the RGB-in/RGB-out interface can preserve coarse scene geometry, while editor backbones differ in texture realism, semantic fidelity, and layout drift.

the DPT-Large RMSE-255 condition score.

Table 11 compares zero-shot RINO with trained controllable-generation systems. Among the zero-shot editors, Qwen gives the best balanced result: RMSE-255 33.83, FID 19.44, and CLIPScore 30.15. FireRed has slightly higher CLIPScore (30.34) but worse depth fidelity, while LongCat keeps reasonable RMSE but falls behind in FID and CLIPScore. Qwen remains behind the strongest trained control models, such as ControlNet++ and ControlAR, which use dedicated conditioning mechanisms. However, it is competitive with several trained adapter/control baselines in depth fidelity while using only a zero-shot RGB editing interface. This supports the central RINO hypothesis: spatial controls can be consumed as images without introducing a separate encoder or adapter for each condition type. Qualitative examples in Figure 10 show that the editors usually respect coarse depth layout, while differing in texture realism and semantic faithfulness.

**Semantic Segmentation Map Conditioned Generation.** Following the ControlNet++ protocol (Li et al., 2024a), each validation sample provides an RGB semantic segmentation map and a text caption. RINO treats the segmentation map as an ordinary RGB conditioning image and instructs the editor to synthesize a natural image whose semantic layout follows the per-region class colors. No segmentation encoder, ControlNet branch, adapter, or task-specific decoder is added: the model only sees an RGB condition image and a natural-language instruction. Because we observe that the richness of the prompt has a large impact on both image quality and condition fidelity, we re-caption every validation image once with an open-source VLM and use the resulting caption as the text prompt for all editors. We evaluate Qwen-Image-Edit (Wu et al., 2025), FireRed-Image-Edit-1.1 (FireRed Team, 2025), and LongCat-Image-Edit (Ma et al., 2025) on the ADE20K (Zhou et al., 2017) and COCOStuff (Lin et al., 2014) validation splits at the native input resolution. Each editor uses its recommended sampling configuration: Qwen-Image-Edit with `steps=40` and `CFG=4.0`, FireRed-Image-Edit-1.1 with `steps=40` and `CFG=4.0`, and LongCat-Image-Edit with `steps=50` and `CFG=4.5`. We follow the ControlNet++ evaluation protocol for segmentation-conditioned generation (Li et al., 2024a). Condition fidelity is measured by re-extracting a semantic segmentation map from each generated image with a Mask2Former (Cheng et al., 2022) segmenter and comparing it against the input condition; we report mean Intersection over Union (mIoU), mean class accuracy (mAcc), and all-pixel accuracy (aAcc). Image quality is measured by FID against the real validation images, and text alignment by CLIPScore against the re-captioned prompt. Lower FID is better; higher mIoU, mAcc, aAcc, and CLIPScore are better.

**Table 12** Semantic segmentation map conditioned image generation on ADE20K and COCOStuff. ControlNet++ is included as a task-trained grounded generation reference. Image-editor rows are run by us zero-shot with RGB segmentation maps and VLM-recaptioned prompts at the native input resolution.

| | ADE20K | | | | | COCOStuff | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | mIoU ↑ | mAcc ↑ | aAcc ↑ | FID ↓ | CLIP ↑ | mIoU ↑ | mAcc ↑ | aAcc ↑ | FID ↓ | CLIP ↑ |
| ControlNet | 32.55 | - | - | 33.28 | 31.53 | 27.46 | - | - | 21.33 | 13.31 |
| ControlNet++ | 43.64 | - | - | 29.49 | 31.96 | 34.56 | - | - | 19.29 | 13.13 |
| *Generative image editors (ours, zero-shot, RGB→RGB)* | | | | | | | | | | |
| Qwen-Image-Edit | 46.24 | 61.92 | 79.13 | 32.05 | 32.62 | 37.40 | 50.64 | 63.50 | 15.54 | 34.14 |
| FireRed-Image-Edit | 45.82 | 62.82 | 78.14 | 31.70 | 32.67 | 37.55 | 51.20 | 63.25 | 15.94 | 34.07 |
| LongCat-Image-Edit | 46.37 | 61.62 | 78.91 | 36.08 | 32.50 | 36.38 | 48.93 | 61.46 | 20.95 | 33.89 |

Table 12 compares the three zero-shot editors against ControlNet++ as a task-trained grounded-generation reference on ADE20K and COCOStuff. All three editors deliver strong condition fidelity: on ADE20K they cluster tightly around 46 mIoU (Qwen 46.24, FireRed 45.82, LongCat 46.37), and on COCOStuff around 37 mIoU (FireRed 37.55, Qwen 37.40, LongCat 36.38). Image quality follows the editor backbone rather than the conditioning interface: Qwen and FireRed are essentially tied on both datasets (31.70–32.05 FID on ADE20K, 15.54–15.94 on COCOStuff), while LongCat trails noticeably (36.08 and 20.95 FID). CLIPScore is also tightly clustered across editors, indicating that the re-captioned prompts are followed comparably well. Overall, a general-purpose image editor consumes a semantic segmentation map as an ordinary RGB condition and respects the per-region class layout zero-shot, with image quality limited by the editor backbone rather than by the lack of a segmentation adapter. Visualizations are shown in Figure 11.


**Figure 11** Qualitative semantic segmentation map conditioned generation on ADE20K. RINO interface preserves the semantic layout of the input, while editor backbones differ in texture realism and small-region fidelity.

**Human-Pose Conditioned Generation.** We adopt the pose-guided text-to-image protocol of HumanSD (Ju et al., 2023b) and Stable-Pose (Wang et al., 2024a) on the Human-Art (Ju et al., 2023a) validation split. For each sample, the ground-truth COCO-17 keypoints are rendered into an OpenPose-style colored skeleton (16 limb segments on a black canvas). The editor receives only this RGB skeleton and the associated caption, and is instructed to align each body part with the corresponding colored bone while omitting the skeleton overlay from the synthesized image. No pose encoder, ControlNet branch, or trainable adapter is introduced. We

**Table 13** Human-pose conditioned image generation on Human-Art validation. Baselines are quoted from Stable-Pose; $*$ marks released checkpoints. Our editor rows are zero-shot. **Bold**: best baseline per column.

| | Pose Accuracy | | Image Quality | | T2I Alignment |
|---|---|---|---|---|---|
| Method | CAP↑ | PCE↓ | FID↓ | KID↓ | CLIP-score↑ |
| *Trained pose-guided generation models* | | | | | |
| SD* (Rombach et al., 2022) | 55.71 | 2.30 | 11.53 | 3.36 | **33.33** |
| T2I-Adapter (Mou et al., 2023) | 65.65 | 1.75 | 11.92 | 2.73 | 33.27 |
| ControlNet (Zhang et al., 2023) | 69.19 | 1.54 | 11.01 | **2.23** | 32.65 |
| Uni-ControlNet (Zhao et al., 2023) | 69.32 | 1.48 | 14.63 | 2.30 | 32.51 |
| GLIGEN (Li et al., 2023b) | 69.15 | 1.46 | – | – | 32.52 |
| HumanSD (Ju et al., 2023b) | 69.68 | **1.37** | **10.03** | 2.70 | 32.24 |
| Stable-Pose (Wang et al., 2024a) | **70.83** | 1.50 | 11.12 | 2.35 | 32.60 |
| *Generative image editors* | | | | | |
| Qwen-Image-Edit | 65.97 | 1.76 | 36.01 | 5.83 | 32.96 |
| FireRed-Image-Edit | 68.09 | 1.82 | 10.66 | 5.23 | 34.24 |
| LongCat-Image-Edit | 55.54 | 2.02 | 7.91 | 6.25 | 32.22 |

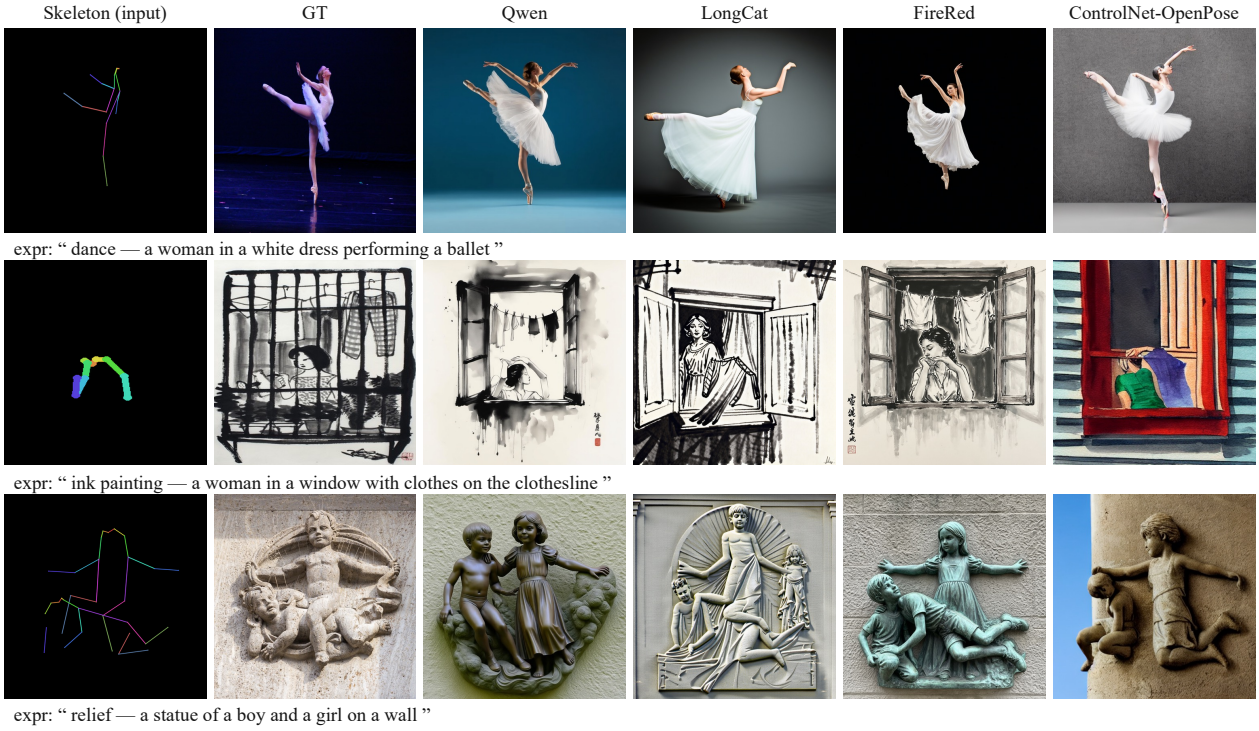


**Figure 12** Zero-shot human-pose conditioned generation across styles: input skeleton, GT, our editors (Qwen/LongCat/FireRed), and ControlNet-OpenPose.

follow the HumanSD/Stable-Pose evaluation protocol. Pose fidelity is assessed by re-estimating the pose of each generated image with HigherHRNet (Cheng et al., 2020) and comparing it to the conditioning skeleton through COCOevalSimilarity, yielding a cosine-similarity-based average precision (CAP), and a people-count error (PCE). Image quality is measured by FID and KID against the real Human-Art images, and text alignment by CLIP-Score. Editor outputs are scored with the public evaluation code under which the baseline numbers were reported.

Table 13 and Figure 12 show that RINO achieves highly competitive zero-shot performance for human-pose conditioned image generation. Although our editors are not trained on this task, their results are already

**Table 14** Instance map conditioned image generation on COCO2017 validation. InstanceDiffusion is included as a task-trained grounded generation reference. Image-editor rows are run by us zero-shot with RGB instance maps and VLM-recaptioned prompts at the native input resolution.

| Method | AP↑ | $AP_{50}$ ↑ | Average Recall↑ | FID↓ |
|---|---|---|---|---|
| *Reported task-trained grounded generation result* | | | | |
| InstanceDiffusion (Wang et al., 2024b) | 27.1 | 50.0 | 38.1 | 25.5 |
| *Generative image editors (ours, zero-shot, RGB instance map input)* | | | | |
| Qwen-Image-Edit (Wu et al., 2025) | 25.4 | 44.3 | 40.8 | 15.3 |
| FireRed-Image-Edit (FireRed Team, 2025) | 23.2 | 42.3 | 37.4 | 16.0 |
| LongCat-Image-Edit (Ma et al., 2025) | 28.0 | 46.3 | 44.2 | 18.5 |

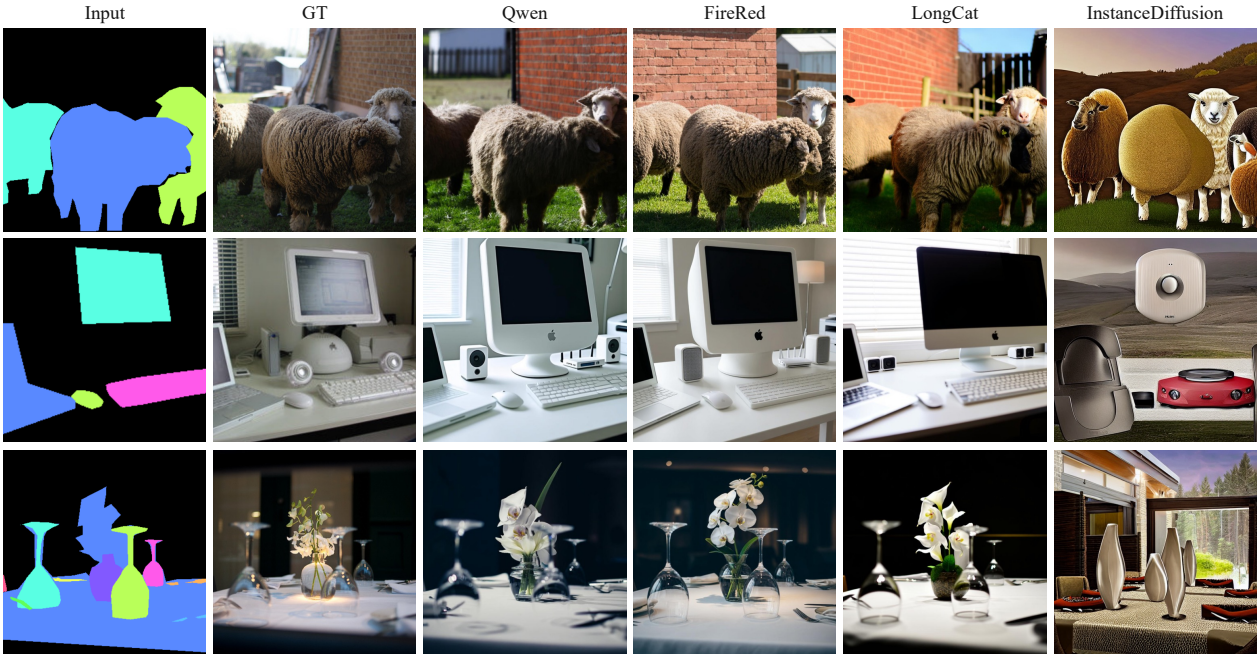


**Figure 13** Qualitative instance map conditioned generation on COCO2017 validation. The same RGB instance maps and re-captioned prompts are given to all three zero-shot editors. The examples show that the RGB-in/RGB-out interface preserves per-instance placements without an instance encoder, while editor backbones differ in texture realism and small-instance fidelity.

comparable to in-domain pose-guided generation models, demonstrating that generic image editors can effectively interpret RGB skeleton inputs as spatial generation conditions. Qualitatively, RINO often produces images that better follow the input pose than ControlNet-based baselines, especially in cases involving unusual body configurations or diverse artistic styles. The generated images also maintain strong text-image alignment and visual realism, suggesting that the RGB-in/RGB-out interface provides an effective and flexible alternative to task-specific pose encoders for conditioned generation.

**Instance Map Conditioned Generation.** Following the InstanceDiffusion protocol (Wang et al., 2024b), each validation sample provides a per-instance map together with a text caption. Unlike a semantic segmentation map, this condition distinguishes individual object instances of the same class rather than just per-category regions. RINO renders the instance map as an ordinary RGB conditioning image and instructs the editor to synthesize a natural image whose object instances follow the input layout. No instance encoder, grounding module, detector head, or trainable adapter is introduced. As in the segmentation-map setting, we re-caption every validation image once with an open-source VLM and use that caption as the prompt for all editors, because we find caption richness materially affects both image quality and the recognizability of generated instances. We evaluate Qwen-Image-Edit, FireRed-Image-Edit, and LongCat-Image-Edit on

the COCO2017 (Wang et al., 2024b) validation split at the native input resolution. Each editor uses its recommended sampling configuration: Qwen-Image-Edit with `steps=40` and `CFG=4.0`, FireRed-Image-Editwith `steps=40` and `CFG=4.0`, and LongCat-Image-Edit with `steps=50` and `CFG=4.5`. We follow the InstanceDiffusion evaluation protocol (Wang et al., 2024b). Condition fidelity is measured by running a pretrained YOLOv8m-Seg (Jocher et al., 2023) detector on the generated images and comparing its detected instance masks against the input condition with COCO's official evaluation metrics. We report mask AP (IoU 0.5:0.05:0.95), $AP_{50}$ (IoU $\geq 0.5$), and Average Recall (AR), which together measure both the localization precision and the recall of the requested instances. Image quality is measured by FID against the real COCO2017 validation images. Higher AP, $AP_{50}$, and AR are better; lower FID is better.

Table 14 compares the three zero-shot editors with InstanceDiffusion as a task-trained grounded-generation reference. The editors follow the per-instance layout closely: LongCat reaches the highest AP (28.0) and AR (44.2), surpassing the trained InstanceDiffusion reference on both metrics, while Qwen (25.4 AP, 44.3 $AP_{50}$, 40.8 AR) and FireRed (23.2 AP, 42.3 $AP_{50}$, 37.4 AR) remain within a few points of it. At the stricter $AP_{50}$ threshold the trained reference still leads (50.0 vs. 42–46 for the editors), indicating that precise instance boundaries are harder to recover than coarse instance placement. Image quality is consistently better than the trained reference across all editors (FID 15.3–18.5 vs. 25.5), reflecting the strength of the modern editor backbones. Overall, a general-purpose image editor can consume a per-instance RGB layout zero-shot and produce images that respect individual instance placements while maintaining higher visual quality than the task-trained baseline. Visualizations are shown in Figure 13.

**Table 15** Canny-conditioned generation on MultiGen-20M validation (5000 images, $512^2$, 4 samples each). Edge controllability is F1 (`cv2.Canny(100,200)`, $\times 100$); image quality is FID; text alignment is CLIP-Score (CLIP ViT-B/16) — all following the ControlNet++ protocol, with FID and CLIP averaged over the 4 generated groups. Our QWEN-IMAGE-EDIT rows are *zero-shot*; trained rows are paper-reported anchors from ControlNet++ (Li et al., 2024a). **Bold**: best zero-shot per column.

| Method | F1 ↑ | FID ↓ | CLIPScore ↑ |
|---|---|---|---|
| *Zero-shot (ours and baselines)* | | | |
| Raw Caption-only | 6.75 | 71.61 | **27.78** |
| Raw Caption + Canny Caption | 7.79 | 59.39 | 23.98 |
| QWEN-IMAGE-EDIT (canny + caption) | **14.90** | **58.30** | 27.31 |
| *Trained for canny conditioning* (Li et al., 2024a) | | | |
| T2I-Adapter (Mou et al., 2023) | 23.65 | 15.96 | 31.71 |
| GLIGEN (Li et al., 2023b) | 26.94 | 18.89 | 31.77 |
| Uni-ControlNet (Zhao et al., 2023) | 27.32 | 17.14 | 31.84 |
| UniControl (Qin et al., 2023) | 30.82 | 19.94 | 31.97 |
| ControlNet (SD1.5) (Zhang et al., 2023) | 34.65 | 14.73 | 32.15 |
| ControlNet++ (SD1.5) (Li et al., 2024a) | 37.04 | 18.23 | 31.87 |

**Canny-Conditioned Image Generation.** Following the MultiGen-20M protocol (Li et al., 2024a), each sample consists of a Canny edge map (white edges on a black background) and a text caption. RINO treats the edge map as an ordinary RGB conditioning image and instructs the editor to synthesize a natural image whose object boundaries and spatial layout conform to the edges, adding no edge encoder, ControlNet branch, adapter, or task-specific decoder. We evaluate Qwen-Image-Edit (Wu et al., 2025) on the 5,000-image validation split at 512×512, drawing 4 samples per condition with `steps=30` and `CFG=4.0` under a shared instruction prompt. All reported numbers are means over the 4 samples per image; no best-of-$N$ selection is used. We follow the ControlNet++ protocol for the Canny condition (Li et al., 2024a). Condition fidelity is measured by re-extracting a Canny edge map from each generated image and comparing it to the input condition; because Canny is a hard (binary) edge, the controllability metric is the per-pixel edge F1-Score, extracted with OpenCV `cv2.Canny(100,200)` to match the protocol exactly. We additionally report image quality as FID against the real validation images and text alignment as CLIP-Score (using CLIP ViT-B/16), both averaged over the 4 generated groups and computed with the same implementations as ControlNet++; the re-extracted edge maps are excluded from the FID image set.

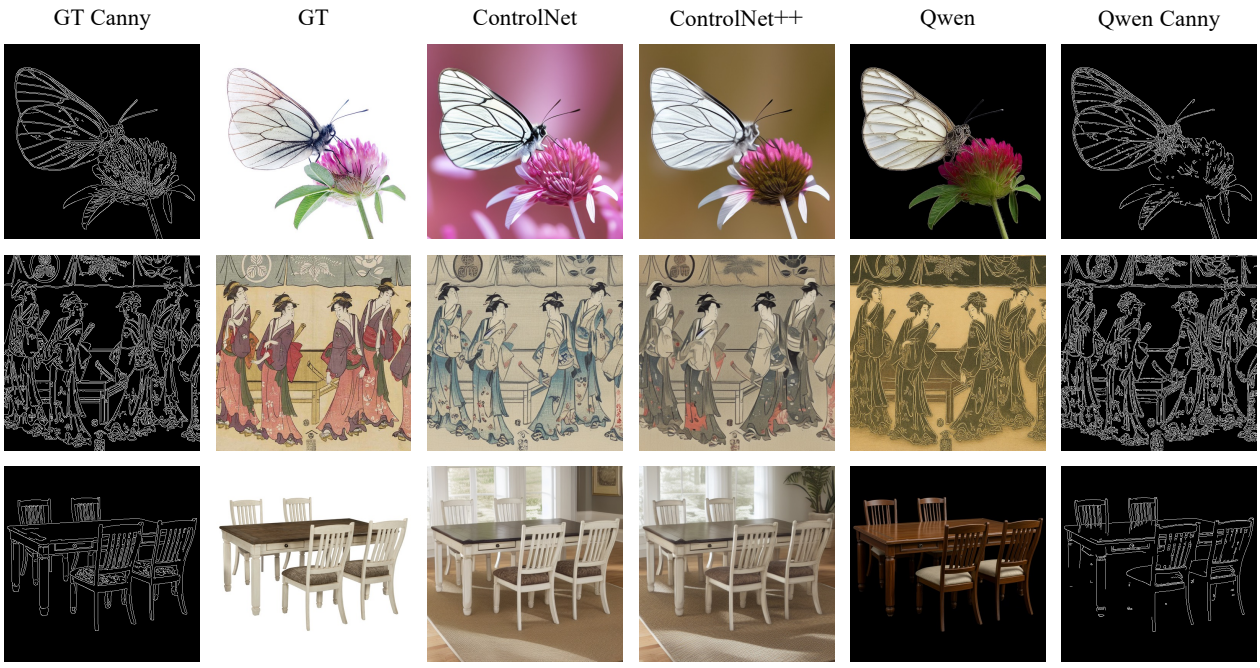


**Figure 14** Qualitative canny-conditioned generation on MultiGen-20M.

Table 15 reports all three metrics. On controllability, zero-shot Qwen-Image-Edit reaches a Canny F1 of 14.90, about 0.40× the purpose-trained ControlNet++ (37.04) and below every trained prior (23–37). The same gap holds on image quality and text alignment: its FID of 59.4 and CLIP-Score of 27.3 trail the trained controllers (FID 15–20, CLIP ≈ 32). Zero-shot canny generation therefore trails purpose-trained controllers on all three axes — edge controllability, distributional realism, and prompt alignment — as expected for a general instruction editor with no edge-conditioned training and no dedicated conditioning branch. The result nonetheless supports the central RINO hypothesis in the harder binary-edge regime: a spatial control can be consumed as an ordinary image, with above-chance controllability, without any task-specific adapter.

To verify that this controllability comes from the Canny *image* rather than the caption, we ablate the condition while holding the editor fixed (top block of Table 15). Removing the edge map and keeping only the caption (*Raw Caption-only*) collapses fidelity to 6.75, and first *describing* the edge map in words with a vision-language model (Qwen2.5-VL-7B) and feeding that text instead of the image (*Raw Caption + Canny Caption*) reaches only 7.79 — both far below the 14.90 from the edge image itself. This implies that the Canny image carries pixel-level spatial information that a language bottleneck cannot substitute. The quality metrics add nuance: supplying structure (the Canny image or its VLM description) improves FID over caption-only (71.6 → 58–59), i.e. a spatial prior yields more realistic images, but routing it through a long VLM description lowers CLIP-Score (27.8 → 24.0), as the appended layout text competes with the caption for prompt alignment. Visualization results are shown in Figure 14.

## 4 Related Work

**Visual estimation models.** Estimation tasks map a natural image to a structured signal, and each task we study has a mature line of specialists that defines its performance ceiling andserves as our reference. Monocular depth estimation progresses from mixing datasets for scale- and shift-invariant relative depth (Ranftl et al., 2022; Birkl et al., 2023; Bhat et al., 2023) to large-scale data-driven and metric-aware systems (Yang et al., 2024a,b; Lin et al., 2025; Wang et al., 2025; Piccinelli et al., 2025), and pretrained diffusion generators are themselves fine-tuned into strong depth and normal estimators (Ke et al., 2024; Ye et al., 2024; He et al., 2025; Bae and Davison, 2024). Segmentation converges on the mask-classification view, where a transformer predicts a set of labeled masks to cover semantic, instance, and panoptic segmentation at once (Cheng et al.,

2021, 2022; Jain et al., 2023; Kerssies et al., 2025; Shahabodini et al., 2025; Li et al., 2023a; Xie et al., 2021), with an open-vocabulary branch that names novel categories via language embeddings (Xu et al., 2022a; Dong et al., 2023; Yu et al., 2023; Barsellotti et al., 2024; Xu et al., 2023). The remaining tasks follow the same pattern of dedicated architectures: pose estimation (Xu et al., 2022b; Maji et al., 2022; Yang et al., 2026b,a), detection and instance segmentation with text-grounded open-set variants (He et al., 2017; Cai and Vasconcelos, 2021; Li et al., 2023a; Liu et al., 2024), and referring segmentation (Yang et al., 2022; Yan et al., 2023; Xiao et al., 2024), many built on general-purpose backbones (Radford et al., 2021; Oquab et al., 2024; Kirillov et al., 2023) but still read out through a task-specific head. Unification here stays *within* a family, with each keeping its own encoder, decoder, and loss.

**Controllable image generation.** Generation tasks move in the opposite direction, synthesizing a natural image from a structured condition—Canny edges, depth, semantic or instance maps, boxes, or poses. ControlNet (Zhang et al., 2023) conditions a text-to-image backbone by attaching a trainable encoder copy that injects the control signal, and later work improves fidelity or extends it to autoregressive backbones (Li et al., 2024a,b); adapter methods instead learn small per-condition side networks (Mou et al., 2023), and unified frameworks route many condition types through a shared module (Zhao et al., 2023; Qin et al., 2023). Layout- and instance-grounded generation (Li et al., 2023b; Wang et al., 2024b; Süleyman and Biricik, 2025) and pose-guided synthesis (Wang et al., 2024a; Ju et al., 2023b) follow the same recipe. The recurring ingredient is a per-condition encoder or adapter trained to inject each new signal. Estimation and generation are thus near-mirror images—one reads a structured signal out of an image, the other writes one in—yet the two traditionally use separate architectures and output spaces and are never unified under a shared interface.

**Toward a unified model for estimation and generation.** Vision Banana (Gabeur et al., 2026) is the first to bridge the two directions: parameterizing the output space of vision tasks as RGB images reframes perception as image generation, and instruction-tuning a single generator makes it competitive with zero-shot specialists such as Segment Anything (Carion et al., 2025) and the Depth Anything series (Yang et al., 2024a,b; Lin et al., 2025), with open-source editors shown to exhibit similar behavior (Liu et al., 2026). This casts RGB as a candidate common interface for both reading and writing images, analogous to text in language models. Vision Banana, however, is built on a proprietary backbone, leaving open how far this interface carries over to open, publicly released models. RINO shares its RGB-as-interface premise and pushes it to this setting: we take open-source image editors as a frozen black box, add only parameter-free RGB conversions, and place estimation and conditioned generation under one RGB-to-RGB formulation across 25 tasks, measuring how closely this fully zero-shot system approaches the specialists above.

## 5 Conclusion

RINO shows that RGB can act as a common interface for both perception and generation: expressing all visual inputs and outputs as RGB lets a single frozen image editor handle a broad range of tasks without any task-specific encoders, decoders, adapters, or auxiliary parameters. Fully zero-shot and without any training, RINO-Zero approaches in-domain specialist models across 25 tasks spanning dense estimation, 3D geometry, and conditioned generation, and offers a zero-shot alternative to adapter-based controllable generation. Its accuracy remains bounded by the visual priors of the editing backbone and by the parameter-free RGB read-out, leaving a gap to specialists on some tasks; lightly instruction-tuning editors on RGB-formatted visual signals, so that the RINO interface is present during training, is a natural way to close it. These findings point to RGB as a shared interface for vision and invite its extension to 3D vision, video, and world models.

## References


Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9535–9545, 2024.

Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3689–3698, 2024.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5386–5395, 2020.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10995–11005, 2023.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

FireRed Team. FireRed-Image-Edit-1.0. Hugging Face model, `FireRedTeam/FireRed-Image-Edit-1.0`, 2025.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T. Barron, Kyle Genova, Nithish Kannen, Sherry Ben, Yandong Li, Mandy Guo, Suhas Yogin, Yiming Gu, Huizhong Chen, Oliver Wang, Saining Xie, Howard Zhou, Kaiming He, Thomas Funkhouser, Jean-Baptiste Alayrac, and Radu Soricut. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.

Jing He, Haodong Li, Mingzhi Sheng, and Ying-Cong Chen. Lotus-2: Advancing geometric dense prediction with powerful image generative model. *arXiv preprint arXiv:2512.01030*, 2025.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *CVPR*, 2023.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, 2023. URL https://github.com/ultralytics/ultralytics.

Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 618–629, 2023a.

Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. HumanSD: A native skeleton-guided diffusion model for human image generation. In *ICCV*, 2023b.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024.

Tommie Kerssies, Niccolò Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your vit is secretly an image segmentation model. In *CVPR*, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of CNN-based single-image depth estimation methods. In *European Conference on Computer Vision (ECCV) Workshops*, pages 331–348. Springer, 2018.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask DINO: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*, 2023a.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. *arXiv preprint arXiv:2404.07987*, 2024a.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22511–22521, 2023b.

Zongming Li, Tianheng Cheng, Shoufa Chen, Peize Sun, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Controlar: Controllable image generation with autoregressive models. *arXiv preprint arXiv:2410.02705*, 2024b.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, 2024.

Wei Liu, Jiaxin Lin, and Rui Chen. Open-source image editing models are zero-shot vision learners. *arXiv preprint arXiv:2605.04566*, 2026.

Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *CVPR Workshops (CVPRW)*, 2022.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, 2016.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision (ECCV)*, pages 792–807. Springer, 2016.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. In *CVPR*, 2025.

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.

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. In *ICML*, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Sajjad Shahabodini, Mobina Mansoori, Farnoush Bayatmakou, Jamshid Abouei, Konstantinos N. Plataniotis, and Arash Mohammadi. The missing point in vision transformers for universal image segmentation. *arXiv preprint arXiv:2505.19795*, 2025.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision (ECCV)*, pages 746–760. Springer, 2012.

Ahmad Süleyman and Göksel Biricik. Grounding text-to-image diffusion models for controlled high-quality image generation. *arXiv preprint arXiv:2501.09194*, 2025.

Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A dense indoor and outdoor DEpth dataset. *arXiv preprint arXiv:1908.00463*, 2019.

Jiajun Wang, Morteza Ghahremani, Yitong Li, Björn Ommer, and Christian Wachinger. Stable-pose: Leveraging transformers for pose-guided text-to-image generation. In *NeurIPS*, 2024a.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6232–6242, 2024b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. OneRef: Unified one-tower expression grounding and segmentation with mask referring modeling. In *NeurIPS*, 2024.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18134–18144, 2022a.

Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2935–2944, 2023.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022b.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:21875–21911, 2024b.

Timing Yang, Sicheng He, Hongyi Jing, Jiawei Yang, Zhijian Liu, Chuhang Zou, and Yue Wang. Fast sam 3d body: Accelerating sam 3d body for real-time full-body human mesh recovery, 2026a. URL https://arxiv.org/abs/2603.15603.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollár, and Kris Kitani. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026b.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18155–18165, 2022.

Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 43(6):1–18, 2024.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36: 32215–32234, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *arXiv preprint arXiv:2305.16322*, 2023.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.