Kiroshi: An Agentic Perception System for High-Accuracy Image Parsing

Haipeng Zhou[1,*], Jinshan Liu[1,3,*], He Zhang[4], Xuequan Lu[5], Jun Ma[1,2], and Lei Zhu[1,2,✉]

[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] The Hong Kong University of Science and Technology
[3] Xi'an Jiaotong University
[4] Adobe
[5] The University of Western Australia

Abstract. Parsing images with high precision for matting and segmentation is significantly more challenging than conventional dense prediction, as it requires accurate estimation of fine-grained details. Existing methods either lack semantic awareness or produce suboptimal predictions; even interactive matting approaches rely on manual verification and repetitive checking, making fully automatic matting still unattainable. In this work, we propose ***Kiroshi***, an agentic perception system for high-accuracy image parsing. We train an Action Model with iterative refinement and mine paired trajectories by sampling grid prompts from residual maps, where each step yields a positive and a negative transition under the same intermediate prediction based on quantitative quality gains. These within-context preference pairs form a reliable supervision signal for post-training to align the MLLM policy toward more effective grid decisions. We also contribute a new High-Fidelity Referring Matting and Segmentation (***HiFiRefMS***) benchmark to evaluate the performance of different models. Experimental results demonstrate that our method surpasses state-of-the-art approaches both quantitatively and qualitatively, and extensive ablation studies further validate the effectiveness and superiority of our agentic design. Project will be released at https://github.com/haipengzhou856/Kiroshi.

Keywords: High-Accuracy Segmentation · Image Matting · MLLM

1 Introduction

Extracting fine-grained, semantically meaningful structures from images is becoming increasingly essential as modern vision systems shift toward high-fidelity scene understanding and content-level manipulation. Applications ranging from AR and VR [49], 3D reconstruction [31], and image editing [71] all demand the ability to recover accurate object boundaries, subtle transparency effects, and instance-specific details. However, despite the progress in high-precision scene parsing, existing solutions still face significant limitations that hinder their applicability in real-world scenarios.

* Co-first authors.
✉ Lei Zhu (leizhu@hkust-gz.edu.cn) is the corresponding author of this work.

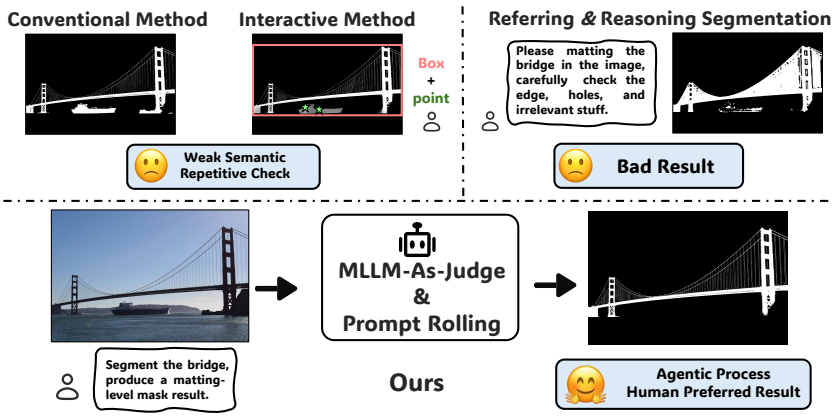

Fig. 1: A comparison of image matting and segmentation paradigms. **(1)** Conventional salient matting methods produce fine details (*e.g.*, ropes) but lack semantic understanding, often failing in multi-instance scenarios. **(2)** Interactive matting relies on human intervention, yet requires repetitive, time-consuming inspection for satisfactory outcomes. **(3)** Referring and reasoning segmentation exhibits semantic comprehension but remains constrained by underlying models (*i.e.*, SAM), leading to coarse mask. Instead, our pipeline establishes an agentic perception loop that autonomously evaluates and refines its predictions, resulting in outputs consistent with human preference.

In prior work on high-accuracy scene parsing as shown in Fig. 1, most methods have been developed for salient-object detection or single-target settings. For instance, in Dichotomous Image Segmentation [42], the dataset provides extremely fine-grained binary annotations but does not distinguish between multiple instances within a scene. In the matting domain, due to the difficulty of obtaining high-quality labels, most datasets primarily focus on limited scenarios, like portraits [25] or animals [26]. Although these datasets have enabled significant progress for training model, a major limitation persists: under the class-agnostic setting, models are highly prone to over-matting or missing regions, especially in complex multi-instance scenarios. Interactive methods [14,17,19,24,65] can partially alleviate this issue. By providing user prompts, the model can extract the desired instance. For example, ZIM [19] scales up SAM [21] using large-scale, high-quality matting data, achieving strong zero-shot matting performance. However, these approaches require frequent user interaction and therefore cannot achieve full automation.

With the rapid rise of multi-model large language models (MLLMs), the vision community has increasingly shifted toward multimodal paradigms. RefMat [27], for instance, introduces the task of referring image matting, where textual descriptions are used to guide the matting process. However, RefMat is entirely built on synthetic cut-and-paste data, resulting in poor generalization to real-world scenarios. Meanwhile, recent MLLM-based reasoning segmentation [13, 22, 36, 75] methods have significantly improved the alignment between visual and textual modalities, yet their prediction quality remains far from fidelity, lacking the ability to recover fine boundaries and transparency effects.

Therefore, despite their strong semantic alignment, these methods still exhibit a substantial gap in segmentation accuracy and practical usage.

To address the limitations of existing approaches, we introduce ***Kiroshi***, an agentic perception system built upon the reasoning capabilities of MLLMs. By jointly leveraging an Action Model and an MLLM-based policy, Kiroshi enables high-precision, fully automated, and semantically aware mask prediction. Concretely, the Action Model performs patch-level refinement by taking the grid map as an attention mask, which provides structured prompting signals and guides the model toward progressively improving the mask. On top of this design, we build a dedicated data curation pipeline and a tailored training strategy that relieve the MLLM from low-level pixel perception; instead, the MLLM is encouraged to acquire patch-level semantic recognition, spatial localization, and error-aware reasoning. We further introduce a trajectory mining strategy within the Action Model, which naturally generates positive–negative trajectory pairs as the prediction evolves across iterations. These grid transitions furnish informative supervision for a subsequent DPO post-training stage, enabling the MLLM to learn autonomous perception, preference alignment, and reliable decision-making in a closed-loop manner. Finally, to validate that Kiroshi effectively addresses the aforementioned challenges, we establish a new high-fidelity benchmark that supports comprehensive and fair comparison against a wide spectrum of *state-of-the-art* baselines.

In summary, in this work the main contributions are:

- We propose an agentic perception system, ***Kiroshi***, for high-accuracy image matting and segmentation. With minimal human instruction or intervention, our framework autonomously reflects on its actions and progressively refines predictions, producing masks.
- To support evaluation, we introduce a new benchmark named ***HiFiRefMS***, and construct a carefully curated training data tailored for our framework.
- Extensive experiments and comparisons demonstrate the effectiveness of our agentic design, highlighting its flexibility, robustness, and scalability in diverse real-world scenarios.

2 Related Works

2.1 High-Accuracy Image Parsing

Reviewing recent progress in the dense prediction community, models have demonstrated continuous improvement in segmentation performance, evolving from CNNs [1,4,45,46] to Transformers [6,16,43,54] and Mamba architectures [34,55,73], as well as transitioning from discriminative [17,21] to generative paradigms [10,56,67]. However, these segmentation approaches remain far from achieving precision, which represents a higher standard of fine-grained accuracy. To this end, several tailored methods have been proposed. Nevertheless, most of them either rely on human intervention [15, 19, 24, 40, 41, 48, 63–65, 69], requiring abundant and complex visual prompts (*e.g.*, trimaps, masks, boxes, points, or scribbles), or are restricted to salient-object scenarios [14,18,42,67,68], leading to suboptimal

matting performance in multi-instance settings. For instance, ZIM [19] inherits the interactive design of SAM [21]. Empowered by large-scale, high-quality matting data, it exhibits strong zero-shot performance but remains insensitive to negative prompts. Moreover, these methods generally lack semantic awareness, resulting in limited capability for context-aware matting.

2.2 Multimodal Scene Parsing

The landscape of multimodal models has expanded rapidly, with emerging variants demonstrating strong capabilities in pixel-level understanding [7]. Early studies primarily focused on referring segmentation [20,30,50,62,70], where textual descriptions are used to prompt segmentation models. For instance, SAM-CLIP [50] leverages these two foundation models for synergistic zero-shot semantic segmentation. With the advancement of LLMs, recent works have increasingly emphasized aligning visual perception with the reasoning ability of LLMs, introducing a new paradigm known as reasoning segmentation. In this schema, LISA [22] employs LoRA [11] to fine-tune the LLM, enabling it to generate a special token <seg> that serves as a prior for SAM's decoder during segmentation. This design has inspired several counterparts [3,44,59,72] exploring the alignment between visual perception and textual semantics. Another interest manner is to treat image patches as decodable tokens [23,47,52], where visual tokens are directly fed into LLMs to generate describable tokens that can be post-processed into segmentation masks. Inspired by this, we construct the visual perception on top of the semantic patch such that our MLLM can yield guidance for prompting the segmentation and matting.

2.3 RLHF in MLLM

The success of DeepSeek-R1 [9] has sparked extensive research on Reinforcement Learning from Human Feedback (RLHF), which uses reinforcement learning to further enhance model capability. This paradigm has also benefited various vision tasks, including image restoration [29], image and video generation [32,57], and dense prediction [13,36,60,74,75]. In the context of segmentation, SAM4MLLM [5] employs an MLLM to localize objects before randomly sampling point prompts, while SegAgent [75] further simulates human annotator trajectories using IoU-based process-supervised rewards [28]. However, reward modeling remains challenging at the pixel-level, and interactive models are often insensitive to point prompts, leading to unstable or random trajectories. Moreover, current MLLMs still struggle with fine-grained pixel-level prediction [7], making it difficult to localize coordinates. These limitations motivate shifting from pixel-wise operations to region-level reasoning, particularly for matting, where much higher precision is required.

3 Approach

3.1 Task Definition and Overview

Given an input image I, the corresponding mask $\mathcal{Y}$, and a textual instruction X , our goal is to predict a high-accuracy mask $\hat{s}$. Depending on the benchmark

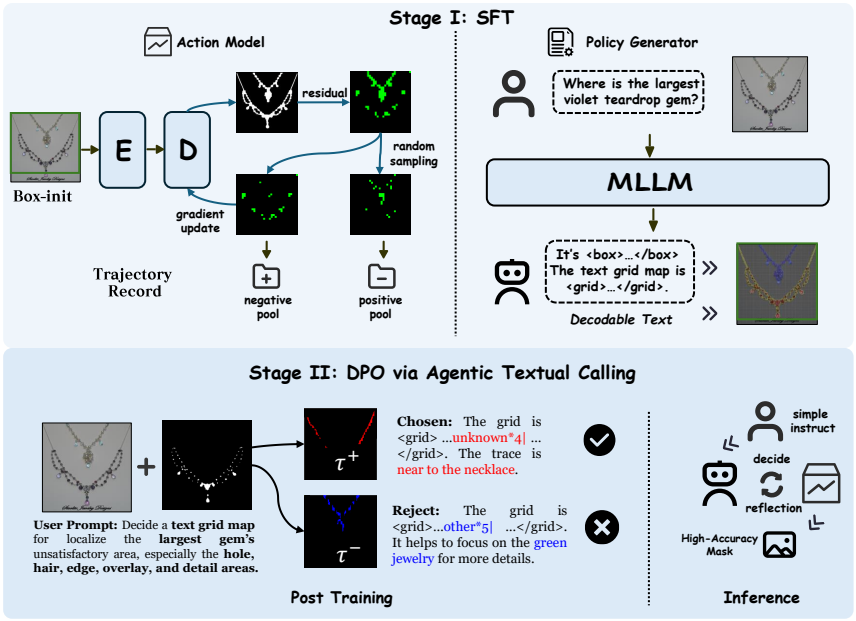

Fig. 2: Overview of the proposed ***Kiroshi*** framework. Our method consists of two training stages. In Stage I, we train an Action Model for mask prediction and fine-tune an MLLM to perceive grid-level visual details. In Stage II, we apply DPO to teach the MLLM to generate more accurate and semantically aligned grid prompts, which then guide the Action Model during inference. With only minimal user intervention, the system is capable of producing high-accuracy masks automatically.

subset, $\hat{s}$ denotes either a binary segmentation mask or a soft matting alpha map, both requiring faithful instance semantics and fine boundary structures.

We view high-accuracy image parsing as an *agentic perception* problem: a system should not only produce an output once, but also *observe its current prediction, decide where it is unreliable, and iteratively refine it*. We propose **Kiroshi**, following an explicit plan-and-act Markov decision process (MDP) with two cooperative components: an MLLM policy generator $\mathcal{P}$ and an action model $\mathcal{A}$. At refinement step t, the policy observes the current state $o_t = (I, X, s_t)$, where s_t is the intermediate parsing result, and emits a region-level action τ_t in the form of a decodable textual grid map:

$$\tau_t = \mathcal{P}(o_t) = \mathcal{P}(I, X, s_t). \tag{1}$$

The action model then executes this action to update the parsing prediction at pixel-level:

$$s_{t+1} = \mathcal{A}(I, s_t, \tau_t), \qquad t = 0, \ldots, T-1. \tag{2}$$

After T steps, the system outputs $\hat{s} = s_T$. Instead of forcing an MLLM coupled with a dense predictor to produce a pixel-accurate mask in a single shot,

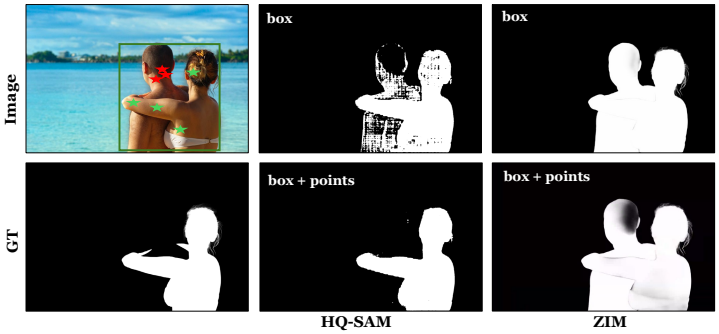

Fig. 3: The human intervention effectiveness test for interactive method.

$\mathcal{P}$ makes instance-aware, region-level decisions by emitting τ_t. These decisions specify ***where* to refine and *what* to refine**, making the refinement process semantically interpretable and the resulting predictions more reliable. Then $\mathcal{A}$ executes the guidance to recover high-frequency details and boundaries, which determines ***how* to refine**. As a result, Kiroshi behaves as an interpretable perception agent that progressively improves its prediction automatically.

3.2 Grid-aware Action Model

We first examine the role of interaction in existing methods such as ZIM [19] and HQ-SAM [17], both of which follow SAM's prompt design [21]. Although negative points are expected to suppress errors, Fig. 3 shows that ZIM, while preserving hair-like structures, often misinterprets negative prompts, whereas HQ-SAM offers more stable semantics but lacks boundary fidelity. Motivated by these limitations, we introduce a grid-aware matting model that incorporates a grid map as an attention mask, enabling structured refinement toward quality. This design **(i)** reduces the search space by replacing unstable pixel-level prompts (*i.e.*, points) with consistent patch-level interaction, **(ii)** provides richer semantic and boundary cues for fine-detail preservation, and **(iii)** naturally aligns with our agentic MLLM framework, facilitating reasoning-driven iterative refinement.

To this end, we design a simple yet effective Action Model equipped with a trajectory mining strategy. The model adopts a DINOV2 [39] encoder and follows the Mask2Former-style decoder [6], where the transformer decoder accepts prompt-generated attention masks to guide the model toward relevant regions and pixel decoder then produces the final refined mask. In detail, the forward process in the layer of transformer decoder could be:

$$\mathbf{Z}_l = \mathrm{softmax}\big(M_{l-1} + \mathbf{Q}_l\mathbf{K}_l^\top\big)\,\mathbf{V}_l + \mathbf{Z}_{l-1}, \tag{3}$$

where $\mathbf{Z}$ is the latent representation. For the attention mask $\mathcal{M}$, we have:

$$\mathcal{M}^b(s) = \begin{cases} 0, & \text{if } (s) \in \text{grid}, \\ -\infty, & \text{otherwise}, \end{cases} \tag{4}$$

Algorithm 1 Trajectory Mining in Action Model $\mathcal{A}$

Require: Image I, label mask $\mathcal{Y}$, initialization box $\mathcal{B}$, rollout steps K, grid size p, threshold ϵ, quality function $Q(\cdot)$
Ensure: Positive pool $\mathcal{D}^+$, Negative pool $\mathcal{D}^-$

Initialize pools $\mathcal{D}^+ \leftarrow \emptyset$, $\mathcal{D}^- \leftarrow \emptyset$
$s_0 \leftarrow \mathcal{A}(I, \mathcal{B})$
for $t = 1$ **to** K **do**
  $r_t \leftarrow |\mathcal{Y} - s_{t-1}|$ ▷ Error Map
  repeat
    $\tau_t^+, \tau_t^- \leftarrow \text{ResidualToGrid}(r_t; p, \epsilon)$ ▷ Two random samples
    $\mathcal{M}_t^+, \mathcal{M}_t^- \leftarrow \text{BuildAttnMask}(\tau_t^+), \text{BuildAttnMask}(\tau_t^-)$
    $s_t^+, s_t^- \leftarrow \mathcal{A}(I, s_{t-1}, \tau_t^+; \mathcal{M}_t^+), \mathcal{A}(I, s_{t-1}, \tau_t^-; \mathcal{M}_t^-)$
    $\Delta_t^+, \Delta_t^- \leftarrow Q(s_t^+) - Q(s_{t-1}), Q(s_t^-) - Q(s_{t-1})$
  until $\Delta_t^+ > 0 \ \wedge \ \Delta_t^- < 0$
  $\mathcal{D}^+ \leftarrow \mathcal{D}^+ \cup \{(I, \mathcal{B}, s_{t-1}, \tau_t^+, s_t^+)\}$
  $\mathcal{D}^- \leftarrow \mathcal{D}^- \cup \{(I, \mathcal{B}, s_{t-1}, \tau_t^-, s_t^-)\}$
  $s_t \leftarrow s_t^+$ ▷ Used for gradient update
end for

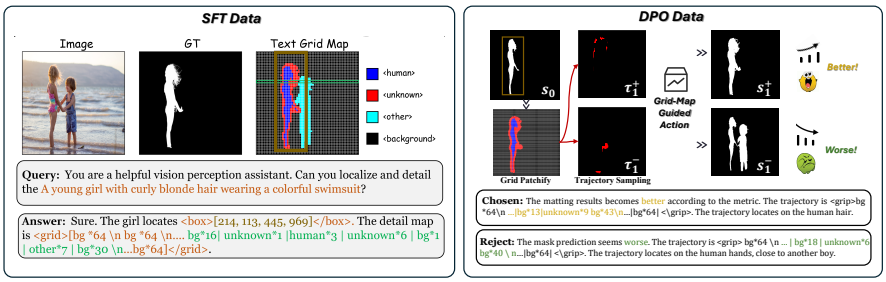

Fig. 4: Data schema for training our MLLM. The left shows SFT data, where the model learns to predict boxes and region-level grid maps from an image–text pair. The right shows DPO data, where masked-guided actions generate alternative trajectories, and preferences are assigned based on matting metrics to guide agentic refinement.

where the grid denotes our prompt areas.

Based on this formulation, we introduce a trajectory mining strategy to prepare within-context preference pairs for post-training (Alg. 1). For each image, we perform K rollouts and collect a pair of grid trajectories (τ_t^+, τ_t^-) at each step. In detail, for a grid patch of size $p \times p$, if more than ϵ (40% by default) of its pixels have non-zero residual values, it is added to the candidate queue for sampling. This is because a zero residual indicates that the model is highly confident in that region and exhibits no uncertainty. We then uniformly sample 75% of the candidate patches to form the trajectory τ. We convert the trajectory grid map into an attention-mask prompt $\mathcal{M}$ and feed it to the action model. We then use quantitative metrics such as MSE to compare the current prediction with the previous one, thereby labeling the transition as positive or negative. As the action model converges, the predicted masks increasingly improve over the previous step, while our random sampling strategy maintains sufficient diver-

sity to still produce informative negative samples. This yields a clean preference signal under the same state $(I, \mathcal{B}, s_{t-1})$ and directly supports DPO optimization.

3.3 Empower MLLM for Perception

Recent MLLM-based approaches [5, 22, 75] can perform referring or reasoning segmentation, yet their predicted masks remain coarse and fail to capture transparency or boundary details. Beyond the inherent limitation of SAM's segmentation granularity, aligning language with ***pixel-level*** dense prediction is intrinsically challenging. When optimization is driven by pixel-level trajectories (*i.e.*, point), the search space becomes extremely large and highly redundant: multiple sampled points within a small 14×14 local region may correspond to nearly identical visual evidence, providing limited additional semantic signal. In contrast, reasoning at the ***patch-level*** abstracts pixels into semantically coherent units, yielding more stable and informative cues. Such region-level alignment allows the probability mass to concentrate on uncertain or erroneous areas, making refinement significantly more tractable for MLLMs.

Therefore, we argue that each image patch can be treated as a textual unit, where a patch serves as a describable token that directly participates in semantic interaction and localization within the LLM. For a given mask, we flatten and parse it into a special textual descriptor, denoted as `<grid>`, where each patch serves as a unit indicating its semantic category. As illustrated in Fig. 4, we define four semantics: the target region (*e.g.*, `<human>`), background (`<bg>`), uncertain areas requiring refinement (`<unknown>`), and other instances (`<other>`). In this manner, the mask can be converted into a purely textual, grid-based descriptor that is directly consumable by the LLM. To reduce token length, we adopt Row-wise Run-Length Encoding (R-RLE) [8, 23] to embed the mask where the consecutive objects are summarized by counts. *E.g.*, if a row contains 64 background patches, we represent it as `bg*64` (single token) instead of 64 separate `bg` tokens. In the worst case, different patch categories may appear in an alternating pattern, yielding the longest possible sequence; for instance, a 32×32 grid over a 1024-resolution image produces up to $32 \times 32 = 1024$ textual tokens. However, due to the inherent spatial continuity of real masks, such extreme fragmentation rarely occurs. As a result, our grid-based textual representation remains highly compact in practice, far shorter than the theoretical upper bound, making the fine-tuning both efficient and lightweight to conduct area grounding. Note that we derive the `<unknown>` region via the trimap using a standard erosion–dilation procedure. Meanwhile, at this stage, we also perform visual grounding to ensure a well-initialized state for subsequent refinement. With this design, the SFT-trained MLLM achieves semantic alignment while acquiring spatial-level parsing capability, enabling it to effectively prompt the Action Model for refinement.

3.4 Alignment in Post-Training

After the previous stage, our MLLM gains semantic awareness and coarse localization ability, but still lacks the capacity to decide ***how*** to refine the mask. Inspired by recent RLHF advances, we adopt Direct Preference Optimization

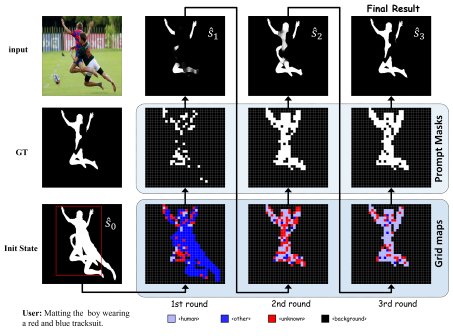

Fig. 5: Inference workflow of our pipeline. Textual response fora the gird map has been decoded and visualized at last row.

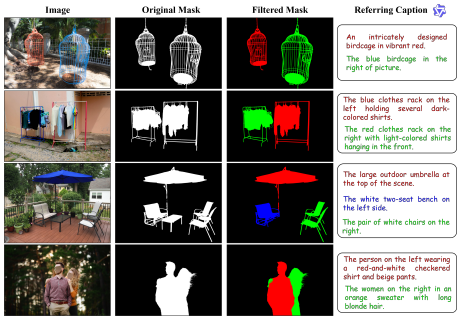

Fig. 6: The data demo of our organized *HiFiRefMS* benchmark. Please zoom in for the best view.

(DPO) to directly align the model with human-preferred results. Since matting involves extremely fine structures and yields sparse pixel-level rewards, DPO offers a more stable and efficient optimization strategy than reinforcement learning methods that rely on dense reward signals.

As illustrated in Fig. 2 and Fig. 4, our DPO data are mined from the Action Model's iterative rollouts. At each refinement step, we compare the updated prediction with the previous one. If the grid prompt τ_t leads to an improvement, the corresponding transition is treated as a preferred sample; otherwise it is marked as rejected. To provide the MLLM with richer semantic and spatial priors, we additionally use Qwen3-VL Plus to generate location-aware, semantically grounded descriptions for each trajectory. These descriptions, together with the chosen/rejected pairs, are organized into VQA-style conversational data to optimize the MLLM with DPO:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\Big(\beta\big[\log \mathcal{P}(\tau_+ \mid I, X) - \log \mathcal{P}(\tau_- \mid I, X)\big]\Big). \tag{5}$$

3.5 Optimization and Inference Schema

In our training pipeline, we optimize the Action Model with a combination of ℓ_1 loss and Laplacian loss [65]. For MLLM fine-tuning, the supervision is purely textual, and we therefore use the standard cross-entropy loss. For the grid map produced by the MLLM, only the `<target>` and `<unknown>` regions are considered valid and are converted into the attention-mask prompt fed to the action model. At inference time, we adopt an MLLM-as-Judge strategy to drive multi-round refinement, as illustrated in Fig. 2 and Fig. 5. Unless otherwise specified, we run 3 refinement loops by default.

4 HiFiRefMS Benchmark

Our goal is to achieve semantically distinguishable, high-precision segmentation and matting with minimal human interaction. However, existing RefCOCO-series datasets [38, 66] provide overly coarse masks, while the Referring Matting

Table 1: Numerical comparison with *state-of-the-art* methods on our *HiFiRefMS* benchmark. We compare our approach with representative methods across multiple categories against salient-based, interactive, and MLLM-based models. The best and runner-up results are highlighted in **red** and **blue**, respectively.

Method	Venue	Segmentation					Matting			
		$maxF_\beta\uparrow$	$F_\beta\uparrow$	MAE↓	$S_\alpha\uparrow$	HCE↓	MSE↓	SAD↓	Grad↓	Conn↓
Salient-based Model										
IS-Net [42]	ECCV'22	0.749	0.693	0.086	0.796	1877	0.1134	138.7	93.44	106.8
PGNet [53]	CVPR'22	0.762	0.719	0.074	0.801	1633	0.1073	149.3	100.7	114.9
HitNet [12]	AAAI'23	0.787	0.730	0.066	0.828	1648	0.1086	122.0	87.46	93.41
MVANet [68]	CVPR'24	0.816	0.762	**0.053**	0.840	1528	0.0635	78.57	63.42	70.40
DiffDIS [67]	NeurIPS'25	**0.829**	**0.807**	0.058	**0.841**	**1386**	0.0601	72.37	68.38	52.88
Interactive Model										
HQ-SAM [17]	NeurIPS'23	0.745	0.729	0.062	0.817	1684	0.0984	83.41	92.68	78.49
MAM [24]	CVPR'24	0.712	0.703	0.073	0.786	1744	0.0597	54.13	48.01	39.42
SmartMat [65]	CVPR'24	0.759	0.736	0.064	0.810	1633	0.0438	42.11	37.47	24.06
SDMatte [14]	ICCV'25	0.766	0.743	0.060	0.807	1629	0.0390	**34.67**	32.48	22.51
ZIM [19]	ICCV'25	0.774	0.767	0.057	0.828	1539	**0.0384**	36.41	**27.88**	**17.49**
MLLM based methods										
LISA [22]	CVPR'24	0.703	0.674	0.106	0.757	2384	0.1208	158.4	113.5	125.7
SAM4MLLM [5]	ECCV'24	0.725	0.698	0.081	0.772	2135	0.0985	143.7	104.3	111.0
Text4Seg [23]	ICLR'25	0.733	0.706	0.074	0.800	2072	0.0844	96.72	94.85	83.62
SegAgent [75]	CVPR'25	0.739	0.718	0.089	0.792	1944	0.0903	87.19	92.33	84.36
Ours	/	**0.875**	**0.820**	**0.033**	**0.866**	**1218**	**0.0263**	**26.34**	**23.71**	**13.48**

dataset [27] relies on synthetic composites, limiting its real-world generalization. Moreover, most the high-quality datasets [25, 26, 35, 42, 48, 69] focus solely on salient objects and lack textual annotations, preventing referential distinction across multiple instances. To this end, we introduce the High-Fidelity Referring Matting and Segmentation (*HiFiRefMS*) benchmark. It consists of two subsets defined by task metrics: matting (468 images / 1187 masks) and segmentation (500 images / 1366 masks). Specifically, segmentation masks are binary with hard labels of 0 and 1, whereas matting masks contain continuous transparency values ranging from 0 to 1. Therefore, different evaluation metrics are required to properly assess their respective prediction quality.

HiFiRefMS reorganizes, cleans, and re-annotates public datasets [25, 26, 35, 42, 48, 69] using real-world images, with an emphasis on multi-instance scenarios. In detail, as shown in Fig. 6, instances and masks in the first two rows can be separated using simple connectivity filtered rules. However, for occluded or visually complex cases in the third and fourth rows, we manually re-annotated the objects to ensure proper instance separation. It is worth noting that we exclude transparent objects from the segmentation subset, as they are treated as foreground in segmentation but require transparency estimation in matting. To avoid such label inconsistency, we directly discard these samples. As illustrated in the third row of Fig. 6, the glass table is removed from the segmentation

annotations. For the textual captions, we employed an API-calling pipeline based on Qwen3-VL Plus [61] to generate referring expressions for each instance.

5 Experiments

5.1 Setups

Implement Details. All experiments are conducted on NVIDIA A100 (80GB) GPUs. For the Action Model, we adopt DINOv2-Base [39] with a mask decoder for parsing. For MLLM, we use Ovis2.5-2B [37] as the backbone. Due to space limitations, more details are provided in the ***Supplementary Material***.

Data Usage. In SFT stage, for the MLLM we adopt publicly available referring data, namely the RefCOCO series [38,66] and RefMatte [27]. We emphasize that, at this stage, the MLLM is trained to acquire semantic perception at the grid-patch level. Therefore, this stage primarily requires sufficient data volume . We also apply dilation and erosion to the masks to obtain a trimap, from which the `<unknown>` region is derived. Such that, a VQA-style conversational format (Fig. 4) is deployed to effectively transfer its object–language grounding ability. For training the action model, we use the public matting dataset comparable in scale to SmartMat [65], and additionally incorporate segmentation datasets DIS-5K [42] and UHR-SOD [53], ensuring the model can handle both segmentation and matting. For DPO post-training, we leverage the trajectory pools described in Sec. 3.2, where the data volume is about 40,000 paired conversations.

5.2 Comparison Configuration

Metric. Consistent with established evaluation practices, we adopt separate sets of metrics for assessing matting and segmentation performance. For image matting, the evaluation is conducted using four commonly employed measures [58]: Sum of Absolute Differences (SAD), Mean Squared Error (MSE), Gradient Distortion (Grad), and Connectivity Loss (Conn). Together, these metrics capture both the overall accuracy and the fine-grained consistency of the estimated alpha mattes. Regarding segmentation, we adhere to the evaluation procedure described in [42], utilizing widely recognized metrics including the maximum F-measure ($maxF_\beta$), weighted F-measure (F_β^w), mean absolute error (M), structural similarity index (S_α), and human correction efforts (HCE).

Compared Methods. We benchmark three categories of representative approaches. **(1)***Salient-based models* [12,42,53,67,68] mainly rely on contrast cues and thus provide limited semantics. **(2)***Interactive models* [14,17,19,24,65] leverage point or box prompts for refinement and achieve better boundary accuracy, yet their quality still depends on repetitive human intervention. **(3)***MLLM-based models* [5,22,23,75] exploit multimodal reasoning to locate objects but remain limited in pixel-level fidelity, leading to inferior matting performance. Together, these methods form complementary baselines for evaluation. Notably, the interactive methods are provided with GT-derived boxes following [17] (*i.e.*, accurate boxes without text referring), except MAM [24] which retains its original referring pipeline: GroundingDINO [33] (text→box)→SAM (mask)→matting. Together, these methods form complementary baselines for evaluation.

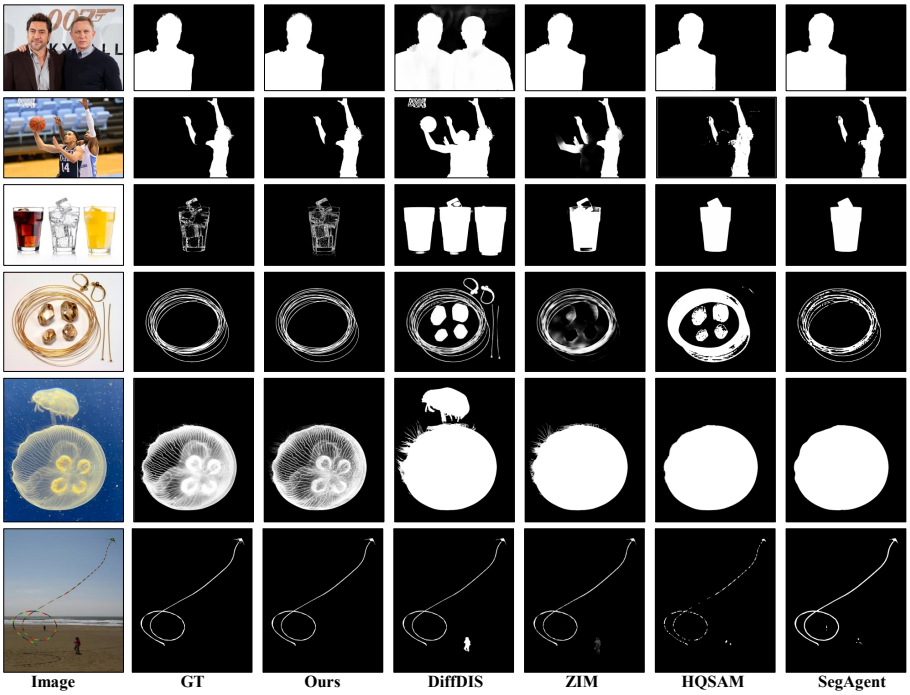

Fig. 7: Visual comparisons with *state-of-the-art* methods. Please zoom in for the best view to observe the hair, transparency, and other details.

5.3 Main Results

Quantitative Comparisons. Tab. 1 shows that our method achieves the best overall performance on *HiFiRefMS*. We obtain the lowest matting errors (MSE/SAD/Grad/Conn) and the strongest segmentation quality ($maxF_\beta$, F_β, MAE, S_α, and HCE), indicating both higher boundary fidelity and better region consistency. While salient-based methods perform reasonably on segmentation-style metrics, they fall behind on matting measures, and MLLM-based methods remain limited in pixel-level precision.

Visual Results. Fig. 7 confirms these trends. Box-driven interactive pipelines can refine boundaries but are inherently ambiguous: a bounding box inevitably contains multiple "stuff" regions or nearby instances, leading to leakage when all contents inside the box are parsed. Referring-based MLLM methods better capture target semantics, yet their masks are often coarse and lose hair, transparency, and thin structures. Our results preserve fine structures with cleaner instance separation.

Comparison Summary. In summary, box-based interaction provides strong low-level cues but suffers from box ambiguity, while referring-only semantics are

Table 2: Performance comparisons of different MLLM backbones.

Backbone	Segmentation			Matting		
	$maxF_\beta \uparrow$	$S_\alpha \uparrow$	HCE↓	SAD↓	Grad↓	Conn↓
Qwen2.5-VL-3B [2]	0.852	0.850	1366	27.51	28.40	15.80
InternVL3.5-2B [51]	0.840	0.822	1478	26.88	25.41	13.99
Qwen3-VL-2B [61]	0.857	0.835	1398	**24.37**	25.13	14.52
Ovis2.5-2B [37]	**0.875**	**0.866**	**1218**	26.34	**23.71**	**13.48**

Table 3: Performance comparison of different policies during inference.

Calling	Segmentation			Matting		
	$maxF_\beta \uparrow$	$S_\alpha \uparrow$	HCE↓	SAD↓	Grad↓	Conn↓
Vanilla (box)	0.764	0.784	1601	32.16	34.84	24.72
1-round	0.838	0.827	1514	28.20	25.41	18.63
3-round	**0.875**	0.866	**1218**	26.34	**23.71**	**13.48**
5-round	0.863	**0.868**	1276	**25.87**	24.10	13.77

reliable yet less precise. Kiroshi combines region-level semantic decisions with iterative pixel-level refinement, reducing ambiguity and improving high-fidelity parsing across both Tab. 1 and Fig. 7.

5.4 Ablation Study

Here, we conduct the ablation study on our *HiFiRefMS* benchmark, on both of segmentation and matting subset.

MLLM Backbone. Tab. 2 indicates that different MLLM backbones lead to noticeable variations on both subsets, suggesting that region-level grounding quality directly affects the downstream refinement. Overall, Ovis2.5-2B achieves the most consistent improvements across the segmentation metrics ($maxF_\beta$, S_α, and HCE) and the matting metrics (SAD, Grad, and Conn), and we therefore use it as the default backbone in all experiments.

Policy Configuration. Tab. 3 studies how the refinement policy and the number of calls influence performance. Compared with the vanilla box-only setting, introducing DPO-based policy optimization yields clear gains on both segmentation and matting, validating the effectiveness of preference-aligned refinement. We also observe a trade-off when increasing the number of refinement calls: too few calls under-refine hard regions, while too many calls bring diminishing returns. In practice, three calls provide the best overall balance between quality and stability, and we adopt 3 loops as the default configuration.

Table 4: Performance comparisons of different grid sizes.

Grid Size	Segmentation			Matting		
	$maxF_\beta \uparrow$	$S_\alpha \uparrow$	HCE↓	SAD↓	Grad↓	Conn↓
16×16	0.862	0.854	1357	26.49	25.43	14.69
32×32	**0.875**	**0.866**	**1218**	**26.34**	**23.71**	**13.48**
64×64	0.820	0.809	1542	29.53	30.02	21.81

Grid Size. Tab. 4 reveals a trade-off between guidance granularity and stability. A finer grid (16×16) provides more detailed spatial cues but can introduce noisier, less consistent trajectories, while a coarser grid (64×64) is more stable yet lacks the resolution needed for precise boundary refinement. Overall, 32×32 offers the best balance across both segmentation and matting, and we use it by default.

5.5 Discussion on Protocol

Here, we further conduct an in-depth comparison and analysis against our counterparts, and discuss the necessity of our agentic system. Tab. 5 presents a cross-validation on both real-world and synthetic dataset. For interactive pipelines

Table 5: More referring comparisons on cross-datasets: our *HiFiRefMS* and synthetic dataset *RefMatte-Syn* [27]. Configs: ① GT-box+ZIM; ② GroundDINO-box+ZIM; ③ Vanilla Text4Seg; ④ Text4Seg-mask+MGMatting; ⑤ Vanilla SegAgent; ⑥ SegAgent-mask+MGMatting; Time is measured per image. We use different numbers of * to indicate the number of models involved and deploy $\times N$ to denote the times of refining using action model (*i.e.*, ZIM, MGMatting, and our $\mathcal{A}$, respectively).

Config	Segmentation			Matting			RefMatte-Syn			Time (s)
	$maxF_\beta \uparrow$	$S_\alpha \uparrow$	HCE↓	SAD↓	Grad↓	Conn↓	SAD↓	Grad↓	Conn↓	
*①×1	0.774	0.828	1539	36.41	27.88	17.49	26.88	24.50	15.47	1.04
**②×1	0.753	0.810	1655	39.83	28.40	17.98	28.43	27.81	19.33	2.98
**③×0	0.706	0.800	2072	96.72	94.85	83.62	69.42	72.30	60.72	7.02
***④×1	0.846	0.818	1603	34.52	25.00	15.34	21.76	20.06	13.27	7.89
***④×3	0.833	0.810	1654	33.92	26.58	15.88	20.87	19.70	14.68	10.43
**⑤×0	0.718	0.792	1944	87.19	92.33	84.36	66.47	69.01	59.75	10.21
***⑥×1	0.834	0.825	1577	35.92	28.44	17.60	23.49	20.84	14.86	11.08
***⑥×3	0.840	0.833	1580	34.54	28.05	15.72	22.67	18.75	14.33	19.48
**Ours×1	0.838	0.827	1514	28.20	25.41	18.63	23.40	21.00	14.56	7.42
**Ours×3	**0.875**	**0.866**	**1218**	**26.34**	**23.71**	**13.48**	**20.18**	**17.43**	**12.72**	12.94

(*e.g.*, ZIM), the method itself only supports visual prompts such as boxes or points. To enable referring semantics, an additional visual grounding model (*i.e.*, GroundDINO) must be introduced to provide a text-conditioned box (Config ②). However, this protocol still cannot avoid the inherent ambiguity of bounding boxes. For referring baselines, the vanilla setting relies on an MLLM with SAM-style masks, and further refinement requires plugging in an external mask-guided module (here MGMatting [69]). Nevertheless , these extended pipelines show limited benefit from additional iterations ($\times 3$), since their policies are not designed for quality-aware reflection (Text4Seg lacks explicit reflection, while SegAgent mainly optimizes SAM via point sampling, and mask-guided refinement is less sensitive to hard boundary cases). Moreover, such a scheme requires triple models resulting in more complexity.

In contrast, our agentic design yields consistent improvements across datasets and iterations: a key reason is that our trajectory is dynamic and determined by the prediction quality. As shown in Fig. 5, the grid map progressively shifts toward boundary and uncertain regions, explicitly prompting the action model to focus on the hardest areas and leading to better results, which validates the necessity and effectiveness of our agentic perception system.

6 Conclusion

In this work, we presented ***Kiroshi***, an agentic perception framework designed to bridge the gap between semantic understanding and precision. By coupling a grid-aware Action Model with an MLLM policy trained through trajectory-based DPO, our system is capable of autonomously reflecting on its predictions and refining them with minimal user intervention. To support both training and evaluation, we further introduced the *HiFiRefMS* benchmark, constructed through

a dedicated data curation pipeline to enable high-fidelity referring matting and segmentation. The extensive experiments demonstrate that *Kiroshi* consistently outperforms *state-of-the-art* methods across both quantitative metrics and visual quality, particularly in recovering fine structures and handling multi-instance scenarios. Our method provides interpretable semantic trajectories during parsing, offering a reliable, effective, and insightful reward design for dense prediction. We believe this work provides a promising direction for unifying reasoning-driven guidance with high-accuracy image parsing.

Acknowledgment. This work is supported by the Guangdong Science and Technology Department (2024ZDZX2004), the Program of Guangdong Education Department, AI Research and Learning Base of Urban Culture under Project 2023WZJD008, and Adobe Research Gift Grant.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE TPAMI **39**(12), 2481–2495 (2017)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bao, X., Sun, S., Ma, S., Zheng, K., Guo, Y., Zhao, G., Zheng, Y., Wang, X.: Cores: Orchestrating the dance of reasoning and segmentation. In: ECCV. pp. 187–204. Springer (2024)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE TPAMI **40**(4), 834–848 (2017)
5. Chen, Y.C., Li, W.H., Sun, C., Wang, Y.C.F., Chen, C.S.: Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In: ECCV. pp. 323–340. Springer (2024)
6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1290–1299 (2022)
7. Ding, H., Tang, S., He, S., Liu, C., Wu, Z., Jiang, Y.G.: Multimodal referring segmentation: A survey. arXiv preprint arXiv:2508.00265 (2025)
8. Golomb, S.: Run-length encodings (corresp.). IEEE transactions on information theory **12**(3), 399–401 (1966)
9. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
10. He, J., Haodong, L., Yin, W., Liang, Y., Li, L., Zhou, K., Zhang, H., Liu, B., Chen, Y.C.: Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In: ICLR (2025)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
12. Hu, X., Wang, S., Qin, X., Dai, H., Ren, W., Luo, D., Tai, Y., Shao, L.: High-resolution iterative feedback network for camouflaged object detection. In: AAAI. vol. 37, pp. 881–889 (2023)
13. Huang, J., Xu, Z., Zhou, J., Liu, T., Xiao, Y., Ou, M., Ji, B., Li, X., Yuan, K.: Sam-r1: Leveraging sam for reward feedback in multimodal segmentation via reinforcement learning. arXiv preprint arXiv:2505.22596 (2025)
14. Huang, L., Liang, Y., Zhang, H., Chen, J., Dong, W., Chen, L., Liu, W., Li, B., Jiang, P.T.: Sdmatte: Grafting diffusion models for interactive matting. arXiv preprint arXiv:2508.00443 (2025)
15. Huynh, C., Oh, S.W., Shrivastava, A., Lee, J.Y.: Maggie: Masked guided gradual human instance matting. In: CVPR. pp. 3870–3879 (2024)
16. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: CVPR. pp. 2989–2998 (2023)
17. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F., et al.: Segment anything in high quality. NeurIPS **36**, 29914–29934 (2023)
18. Ke, Z., Sun, J., Li, K., Yan, Q., Lau, R.W.: Modnet: Real-time trimap-free portrait matting via objective decomposition. In: AAAI. vol. 36, pp. 1140–1147 (2022)

19. Kim, B., Shin, C., Jeong, J., Jung, H., Lee, S.Y., Chun, S., Hwang, D.H., Yu, J.: Zim: Zero-shot image matting for anything. In: ICCV. pp. 23828–23838 (2025)
20. Kim, D., Kim, N., Lan, C., Kwak, S.: Shatter and gather: Learning referring image segmentation with text supervision. In: ICCV. pp. 15547–15557 (2023)
21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
22. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR. pp. 9579–9589 (2024)
23. Lan, M., Chen, C., Zhou, Y., Xu, J., Ke, Y., Wang, X., Feng, L., Zhang, W.: Text4seg: Reimagining image segmentation as text generation. In: ICLR (2025)
24. Li, J., Jain, J., Shi, H.: Matting anything. In: CVPR. pp. 1775–1785 (2024)
25. Li, J., Ma, S., Zhang, J., Tao, D.: Privacy-preserving portrait matting. In: ACM MM. pp. 3501–3509 (2021)
26. Li, J., Zhang, J., Maybank, S.J., Tao, D.: Bridging composite and real: towards end-to-end deep image matting. IJCV **130**(2), 246–266 (2022)
27. Li, J., Zhang, J., Tao, D.: Referring image matting. In: CVPR. pp. 22448–22457 (2023)
28. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let's verify step by step. In: ICLR (2023)
29. Lin, Y., Lin, Z., Chen, H., Pan, P., Li, C., Chen, S., Wen, K., Jin, Y., Li, W., Ding, X.: Jarvisir: Elevating autonomous driving perception with intelligent image restoration. In: CVPR. pp. 22369–22380 (2025)
30. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: CVPR. pp. 23592–23601 (2023)
31. Liu, F., Tran, L., Liu, X.: Fully understanding generic objects: Modeling, segmentation, and reconstruction. In: CVPR. pp. 7423–7433 (2021)
32. Liu, R., Wu, H., Zheng, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. In: CVPR. pp. 8009–8019 (2025)
33. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV. pp. 38–55. Springer (2024)
34. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. NeurIPS **37**, 103031–103063 (2024)
35. Liu, Y., Xie, J., Shi, X., Qiao, Y., Huang, Y., Tang, Y., Yang, X.: Tripartite information mining and integration for image matting. In: ICCV. pp. 7555–7564 (2021)
36. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)
37. Lu, S., Li, Y., Xia, Y., Hu, Y., Zhao, S., Ma, Y., Wei, Z., Li, Y., Duan, L., Zhao, J., et al.: Ovis2. 5 technical report. arXiv preprint arXiv:2508.11737 (2025)
38. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR. pp. 11–20 (2016)
39. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
40. Park, G., Son, S., Yoo, J., Kim, S., Kwak, N.: Matteformer: Transformer-based image matting via prior-tokens. In: CVPR. pp. 11696–11706 (2022)

41. Park, K., Woo, S., Oh, S.W., Kweon, I.S., Lee, J.Y.: Mask-guided matting in the wild. In: CVPR. pp. 1992–2001 (2023)
42. Qin, X., Dai, H., Hu, X., Fan, D.P., Shao, L., Van Gool, L.: Highly accurate dichotomous image segmentation. In: ECCV. pp. 38–56. Springer (2022)
43. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021)
44. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model. In: CVPR. pp. 26374–26383 (2024)
45. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
46. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE TPAMI **39**(4), 640–651 (2016)
47. Su, Y., Zhang, H., Li, S., Liu, N., Liao, J., Pan, J., Liu, Y., Xing, X., Sun, C., Li, C., et al.: Patch-as-decodable-token: Towards unified multi-modal vision tasks in mllms. arXiv preprint arXiv:2510.01954 (2025)
48. Sun, Y., Tang, C.K., Tai, Y.W.: Human instance matting via mutual guidance and multi-instance refinement. In: CVPR. pp. 2647–2656 (2022)
49. Tian, Y., Bai, H., Zhao, S., Fu, C.W., Yu, C., Qin, H., Wang, Q., Heng, P.A.: Kineappendage: Enhancing freehand vr interaction through transformations of virtual appendages. IEEE TVCG (2022)
50. Wang, H., Vasu, P.K.A., Faghri, F., Vemulapalli, R., Farajtabar, M., Mehta, S., Rastegari, M., Tuzel, O., Pouransari, H.: Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In: CVPR. pp. 3635–3647 (2024)
51. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
52. Wu, S., Jin, S., Zhang, W., Xu, L., Liu, W., Li, W., Loy, C.C.: F-lmm: Grounding frozen large multimodal models. In: CVPR. pp. 24710–24721 (2025)
53. Xie, C., Xia, C., Ma, M., Zhao, Z., Chen, X., Li, J.: Pyramid grafting network for one-stage high resolution saliency detection. In: CVPR. pp. 11717–11726 (2022)
54. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. NeurIPS **34**, 12077–12090 (2021)
55. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In: MICCAI. pp. 578–588. Springer (2024)
56. Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., Shen, C.: What matters when repurposing diffusion models for general dense perception tasks? In: ICLR (2024)
57. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. NeurIPS **36**, 15903–15935 (2023)
58. Xu, N., Price, B., Cohen, S., Huang, T.: Deep image matting. In: CVPR. pp. 2970–2979 (2017)
59. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: Visa: Reasoning video object segmentation via large language models. In: ECCV. pp. 98–115. Springer (2024)
60. Yan, Z., Li, Z., He, Y., Wang, C., Li, K., Li, X., Zeng, X., Wang, Z., Wang, Y., Qiao, Y., et al.: Task preference optimization: Improving multimodal large language models with vision task alignment. In: CVPR. pp. 29880–29892 (2025)

61. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
62. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: CVPR. pp. 18155–18165 (2022)
63. Yao, J., Wang, X., Yang, S., Wang, B.: Vitmatte: Boosting image matting with pre-trained plain vision transformers. Information Fusion **103**, 102091 (2024)
64. Yao, J., Wang, X., Ye, L., Liu, W.: Matte anything: Interactive natural image matting with segment anything model. Image and Vision Computing **147**, 105067 (2024)
65. Ye, Z., Liu, W., Guo, H., Liang, Y., Hong, C., Lu, H., Cao, Z.: Unifying automatic and interactive matting with pretrained vits. In: CVPR (2024)
66. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV. pp. 69–85. Springer (2016)
67. Yu, Q., Jiang, P.T., Zhang, H., Chen, J., Li, B., Zhang, L., Lu, H.: High-precision dichotomous image segmentation via probing diffusion capacity. In: ICLR (2024)
68. Yu, Q., Zhao, X., Pang, Y., Zhang, L., Lu, H.: Multi-view aggregation network for dichotomous image segmentation. In: CVPR. pp. 3921–3930 (2024)
69. Yu, Q., Zhang, J., Zhang, H., Wang, Y., Lin, Z., Xu, N., Bai, Y., Yuille, A.: Mask guided matting via progressive refinement network. In: CVPR. pp. 1154–1163 (2021)
70. Yu, S., Seo, P.H., Son, J.: Zero-shot referring image segmentation with global-local context features. In: CVPR. pp. 19456–19465 (2023)
71. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
72. Zhu, L., Chen, T., Xu, Q., Liu, X., Ji, D., Wu, H., Soh, D.W., Liu, J.: Popen: Preference-based optimization and ensemble for lvlm-based reasoning segmentation. In: CVPR. pp. 30231–30240 (2025)
73. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: efficient visual representation learning with bidirectional state space model. In: ICML. pp. 62429–62442 (2024)
74. Zhu, L., Ouyang, B., Zhang, Y., Cheng, T., Hu, R., Shen, H., Ran, L., Chen, X., Yu, L., Liu, W., et al.: Lens: Learning to segment anything with unified reinforced reasoning. arXiv preprint arXiv:2508.14153 (2025)
75. Zhu, M., Tian, Y., Chen, H., Zhou, C., Guo, Q., Liu, Y., Yang, M., Shen, C.: Segagent: Exploring pixel understanding capabilities in mllms by imitating human annotator trajectories. In: CVPR. pp. 3686–3696 (2025)