UNIOSE: SEMANTICS-ENRICHED OBJECT QUERIES FOR OPEN-ENDED DETECTION AND SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-ended object detection and segmentation require discovering objects and generating their category names without a predefined category vocabulary. Existing methods typically generate names from object representations produced by a detector. We find that insufficient category semantics in these representations limit naming accuracy, while enriched object semantics can further improve detection and segmentation. We therefore propose UniOSE, a unified framework that constructs semantics-enriched object queries to support both category generation and spatial prediction. Our Semantics-Enriched Object Namer combines detector features with region-level semantics from a pretrained vision–language model, providing richer semantic representations for category generation. Building on these representations, the Semantics-Guided Object Detector incorporates semantic compatibility between object queries and target categories into training-time matching, allowing supervision assignment to account for both object semantics and spatial location. The Semantically Conditioned Mask Decoder injects the same queries into mask decoding, enabling object semantics and local visual evidence to jointly inform mask prediction. Experiments demonstrate strong performance in open-ended object detection and segmentation, including a 5.0 AP gain over the prior method on LVIS minival.

1 INTRODUCTION

Object detection and segmentation are fundamental visual recognition tasks that require identifying objects and determining their spatial extent. Conventional methods are typically trained and evaluated within a fixed category vocabulary, restricting recognition to concepts specified by the annotations (He et al., 2017; Carion et al., 2020; Cheng et al., 2022). Large-scale image–text pretraining offers a way to move beyond this restriction by learning transferable correspondences between visual content and natural language (Radford et al., 2021; Li et al., 2022b). Building on this progress, methods such as ViLD, OVSeg, and OpenSeeD introduce language supervision into region recognition, mask classification, and unified detection and segmentation (Gu et al., 2022; Liang et al., 2023; Zhang et al., 2023b). These open-vocabulary approaches expand the range of recognizable categories, but generally still require candidate category names at inference. When the objects present in an image are unknown beforehand, specifying a comprehensive vocabulary is challenging.

Open-ended perception removes this requirement by asking models to discover objects and generate their names directly from images. Recent studies approach this setting from several directions. GenerateU combines region detection with language generation, Open-Det develops efficient alignment between detectors and language models, and DINO-X integrates object-level recognition with multiple spatial prediction capabilities (Lin et al., 2024a; Cao et al., 2025; Ren et al., 2024). For segmentation, OSM generates names for mask proposals, VL-SAM uses vision–language attention to guide object segmentation, and AutoSeg constructs image-specific vocabularies for semantic segmentation (Yu et al., 2024; Lin et al., 2024b; Ülger et al., 2025). Together, these advances extend recognition beyond supplied category lists and place greater demands on object representations. The model must recover and express an object's category from visual evidence rather than only compare it against given category names.

Despite this progress, existing methods have yet to fully exploit object semantics for both category generation and spatial prediction. As illustrated in Figure 1(a), a representative approach first obtains

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

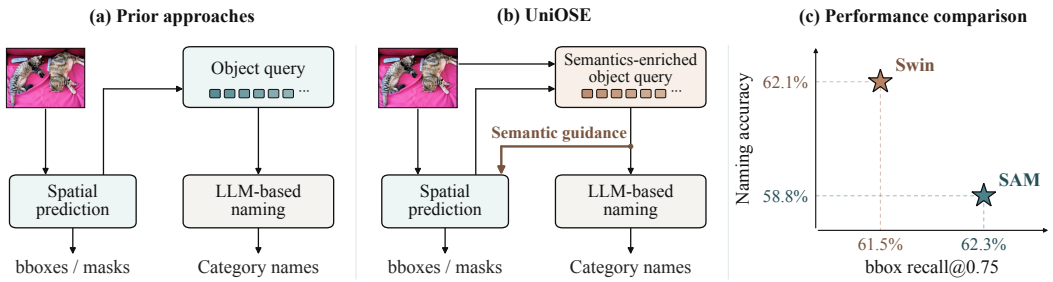


Figure 1: Object semantics for open-ended naming and spatial prediction. (a) Spatial prediction produces object queries for LLM-based naming. (b) UniOSE enriches these queries with image-derived regional semantics and uses them for training-time detection matching and mask conditioning. Semantic guidance originates from the enriched query before LLM-based naming. (c) Comparison of bbox recall@0.75 and naming accuracy. Swin and SAM denote Open-Det Swin-T and its fully fine-tuned SAM2-T backbone variant, respectively. The protocol, metric definitions, and source results are provided in Appendix A.1.

object queries through spatial prediction and then feeds them into a language model to generate category names. However, more accurate spatial prediction does not directly translate into more accurate naming. In Figure 1(c), the baseline with a SAM backbone achieves higher bounding-box recall at an IoU threshold of 0.75 but lower naming accuracy than its Swin counterpart (Appendix A.1), indicating that stronger localization cannot substitute for learning category semantics. Methods such as Open-Det introduce semantic supervision through text alignment (Cao et al., 2025), but the visual evidence for naming still comes primarily from the detector, requiring these features to support both spatial prediction and category generation. Moreover, object category semantics not only underpin accurate naming but also provide complementary cues for localization and region delineation. In particular, when objects are adjacent, occluded, or visually similar, category semantics can help distinguish the objects and their respective regions. However, in existing methods, the semantic information learned for naming is used primarily for category generation and remains underused in spatial prediction. **This motivates us to explore how to enrich object queries with category semantics for more accurate naming and further use these semantics to guide unified spatial prediction.**

To this end, we propose UniOSE, which connects open-ended category generation and spatial prediction through semantics-enriched object queries. As shown in Figure 1(b), UniOSE addresses two aspects of this problem, enriching the category semantics of object queries and using those semantics to guide spatial prediction. First, we design a Semantics-Enriched Object Namer that combines detector features with region-level semantics from a pretrained vision–language model (VLM) to construct a semantics-enriched query for each object candidate. The query then conditions a generative language model to produce the category name. Detector features capture candidate-specific visual information, complemented by regional category semantics. Language alignment and name generation supervision further shape these queries into category-aware representations for open-ended naming.

UniOSE further uses the same semantics-enriched queries for detection and segmentation. The Semantics-Guided Object Detector incorporates semantic compatibility between queries and target categories into training-time matching, allowing supervision assignment to account for both object semantics and spatial location. The Semantically Conditioned Mask Decoder injects the semantics-enriched query into a mask token and combines it with the location cues provided by the candidate box and local visual features to predict the object's spatial extent. Through these two mechanisms, category information in the semantics-enriched queries supports name generation, detection supervision assignment, and mask prediction. Further benchmark evaluations show that UniOSE outperforms existing methods in both open-ended object detection and segmentation, demonstrating the benefits of semantics-enriched queries for both category recognition and spatial prediction. Our main contributions are as follows.

- We propose UniOSE, which connects open-ended category generation and spatial prediction through semantics-enriched object queries. Its Semantics-Enriched Object Namer

combines detector features with region-level semantics from a pretrained VLM to construct richer object representations for category-name generation.

- We design a Semantics-Guided Object Detector and a Semantically Conditioned Mask Decoder that incorporate semantics-enriched queries into detection matching during training and mask prediction, respectively, enabling the same object representation to support category generation, detection, and segmentation.
- Extensive experiments demonstrate that UniOSE with SAM2-T surpasses Swin-T Open-Det by **5.0 AP** on LVIS minival and achieves leading open-ended segmentation results on A-847, PC-459, A-150, and PC-59.

2 Related Work

Open-Vocabulary and Open-Ended Perception. OVR-CNN, ViLD, and Detic extend detection through caption supervision, vision–language distillation, and image-level labels, respectively (Zareian et al., 2021; Gu et al., 2022; Zhou et al., 2022). GLIP and Grounding DINO learn from grounding data (Li et al., 2022c; Liu et al., 2024), while OWL-ViT and OWLv2 explore contrastive pretraining and scalable self-training (Minderer et al., 2022; 2023). For segmentation, LSeg aligns pixel embeddings with text, and CAT-Seg aggregates image–text similarity volumes (Li et al., 2022a; Cho et al., 2024). OVSeg adapts CLIP to masked inputs, MaskCLIP introduces mask tokens into CLIP, SAN combines a side network with frozen CLIP features, and FC-CLIP shares a frozen backbone for mask generation and classification (Liang et al., 2023; Ding et al., 2023; Xu et al., 2023; Yu et al., 2023). Unlike these vocabulary-dependent settings, open-ended methods generate names for discovered objects. GenerateU, Open-Det, and DINO-X name object candidates (Lin et al., 2024a; Cao et al., 2025; Ren et al., 2024), while ZeroSeg and OSM name segmented regions (Rewatbowornwong et al., 2023; Yu et al., 2024). AutoSeg constructs image-specific vocabularies from local captions, whereas VL-SAM converts attention associated with generated names into spatial prompts (Ülger et al., 2025; Lin et al., 2024b). Grounding DINO and Open-Det derive sigmoid–focal matching costs from decoder-query/text alignment scores; OV-DETR conditions queries on CLIP-derived image or text embeddings and uses binary matching to determine whether predictions correspond to the queried concept (Liu et al., 2024; Cao et al., 2025; Zang et al., 2022). UniOSE instead draws on semantics-enriched naming queries to provide the semantic evidence for detection matching.

Region-Level Vision–Language Modeling. RegionCLIP pretrains region–text correspondence (Zhong et al., 2022), while BLIP-2 connects frozen image encoders and language models through learned queries (Li et al., 2023). RegionGPT, Ferret, and Osprey incorporate region-specific visual representations into language models (Guo et al., 2024; You et al., 2024; Yuan et al., 2024); FlexCap conditions descriptions on target boxes and requested caption lengths (Dwibedi et al., 2024). OPAL further supplements generative training with region–text contrastive learning (González et al., 2026). For detector-driven naming, Open-Det aggregates backbone and encoder features with detector queries and learns visual prompts through text-embedding distillation, while DINO-X combines queries with RoIAlign-extracted backbone features (Cao et al., 2025; Ren et al., 2024). OSM uses mask and context queries to aggregate regional content and surrounding information (Yu et al., 2024). UniOSE complements detector-derived features with candidate-aligned region semantics from a separate pretrained VLM, then uses the enriched queries for naming and spatial prediction.

Unified and Promptable Segmentation. OpenSeeD jointly learns detection and segmentation in a common semantic space, while X-Decoder unifies pixel and language outputs through generalized decoding (Zhang et al., 2023b; Zou et al., 2023). SAM and SAM 2 predict masks from spatial prompts (Kirillov et al., 2023; Ravi et al., 2025). This differs from language-conditioned decoding, where semantic representations help specify the target of mask prediction. LISA conditions mask decoding on the language-model state associated with a segmentation token (Lai et al., 2024), while TAP jointly decodes mask and semantic tokens for segmentation, recognition, and captioning (Pan et al., 2024). UniOSE supplies the mask decoder with the enriched object query used for autonomous naming, rather than a language-model state or a re-encoded category name. This query is constructed before name generation and also guides training-time detection matching, connecting the three predictions through the same object-level semantic evidence.

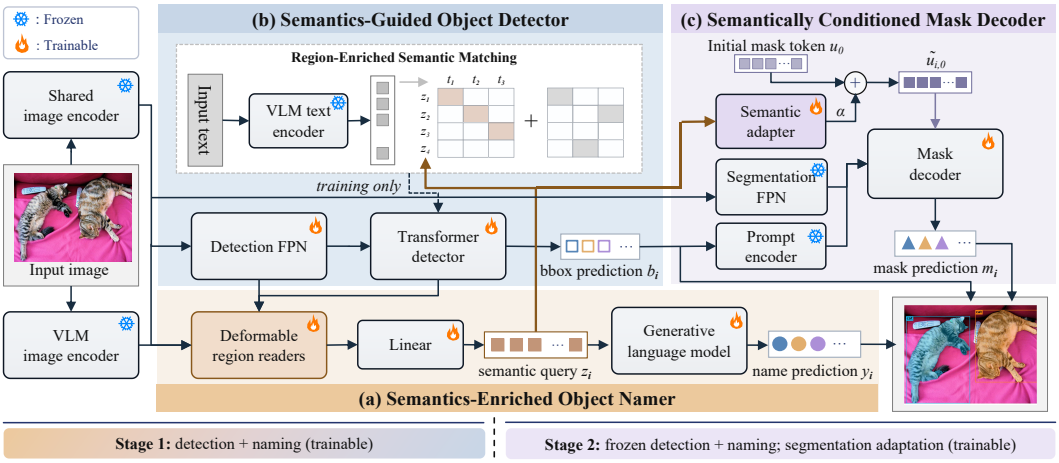


Figure 2: Overview of the UniOSE model. (a) **Semantics-Enriched Object Namer** combines detector features with region-level VLM semantics to construct queries z_i and generates category names with a generative language model. (b) **Semantics-Guided Object Detector** predicts candidate boxes and combines query–target category matching costs with spatial costs for training-time supervision assignment. (c) **Semantically Conditioned Mask Decoder** injects z_i into the first mask token and combines it with image features and predicted boxes to generate masks.

3 METHOD

As shown in Figure 2, given an image $I \in \mathbb{R}^{H\times W\times 3}$, a shared image encoder extracts visual features F^{vis}, which are adapted into detection features F^{det} and segmentation features F^{seg}. A separate pretrained VLM image encoder provides semantic features F^{vl}. The detector processes F^{det} to produce N candidates $\{(q_i, b_i)\}_{i=1}^{N}$, where $q_i \in \mathbb{R}^{d_q}$ is an object query and $b_i \in \mathbb{R}^4$ is its predicted box. The object namer combines detector features with candidate-level VLM features to construct a semantics-enriched object query z_i, hereafter referred to as a semantic query, and conditions a generative language model on it to produce the category-name sequence y_i. During training, the Semantics-Guided Object Detector uses the category semantics in z_i to guide matching. For retained candidates, the Semantically Conditioned Mask Decoder combines z_i, b_i, and F^{seg} to predict a mask m_i. The final output is $\mathcal{O} = \{(b_i, y_i, m_i) : i \in \mathcal{S}\}$, where $\mathcal{S} \subseteq \{1, \ldots, N\}$ indexes the retained candidates. Inference requires only an image, without a category list.

3.1 SEMANTICS-ENRICHED OBJECT NAMER

Existing methods (Lin et al., 2024a; Cao et al., 2025; Ren et al., 2024) primarily rely on detector features to generate category names, and insufficient category semantics in these features limit naming accuracy. To address this issue, we design the **Semantics-Enriched Object Namer**, which combines detector-derived object and contextual features with regional semantics from a pretrained VLM to provide additional evidence for category recognition. The module aggregates multi-source features around the same candidate to construct a semantic query z_i, which conditions a generative language model to generate the category name, as shown in Figure 2(a).

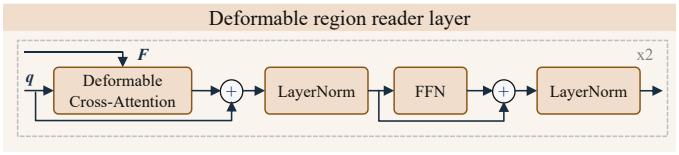


Figure 3: **Structure of a deformable region reader layer.** Each region-reading branch stacks two such layers with branch-specific parameters. The first layer takes q_i as its query; the second takes the updated query. F denotes the branch feature F^s, where $s \in \{\text{enc}, \text{det}, \text{vl}\}$, with channel projection for the VLM-feature branch. $\oplus$ denotes residual addition; the box reference b_i is omitted for clarity.

Specifically, for a candidate (q_i, b_i), three region-reading branches read $F^s \in \{F^{\mathrm{enc}}, F^{\mathrm{det}}, F^{\mathrm{vl}}\}$, corresponding to the detector's encoder memory, detection features, and VLM image features, respectively. The encoder memory is obtained from F^{det}, and the VLM image features are projected to the reader's channel dimension before reading. Each branch starts from q_i and stacks two reader layers with separate parameters across branches. As shown in Figure 3, in each layer, the current query predicts sampling offsets and attention weights for deformable cross-attention (Zhu et al., 2021), using b_i as the spatial reference to aggregate object and surrounding contextual features. The attention update is added to the input query and normalized; an FFN then refines the result, followed by a second residual addition and layer normalization.

The output of the first layer becomes the query input to the second, while both layers read the same branch feature and use the same box reference.

Each branch outputs a region-enriched query r_i^s, where $s \in \{\mathrm{enc}, \mathrm{det}, \mathrm{vl}\}$, aggregating evidence from its own feature source. Separate MLPs map the original object query q_i and these three region-enriched queries to a common semantic space, producing h_i^q and h_i^s, respectively. The projected representations are then concatenated and linearly fused into the semantic query z_i,

$$h_i^q = \mathrm{MLP}_q(q_i), \quad h_i^s = \mathrm{MLP}_s(r_i^s), \quad z_i = \mathrm{Linear}\big([h_i^q; h_i^{\mathrm{enc}}; h_i^{\mathrm{det}}; h_i^{\mathrm{vl}}]\big), \tag{1}$$

where $h_i^q, h_i^s \in \mathbb{R}^{d_s}$, d_s is the shared semantic dimension, $[\cdot]$ denotes channel-wise concatenation, and the linear projection maps the concatenated representation to d_s dimensions. The name generator receives the ℓ_2-normalized z_i as a single visual input token and autoregressively generates the category-name sequence y_i. The namer is trained jointly with the detector using the inherited language-alignment, distillation, and name-generation objectives (Section 3.4); direct semantic-alignment and naming supervision is applied at the final two detector decoder layers. The resulting z_i, before the naming-specific normalization, is also provided to the object detector for semantic matching (Section 3.2) and to the mask decoder for semantic conditioning (Section 3.3).

3.2 Semantics-Guided Object Detector

Following the query-based detection paradigm of DETR (Carion et al., 2020), we adopt the Transformer encoder–decoder architecture of DINO (Zhang et al., 2023a), with the dynamic candidate selection and decoder organization inherited from Open-Det (Cao et al., 2025). The encoder aggregates multi-scale detection features F^{det} into memory F^{enc}, and the decoder refines object queries to predict bounding boxes and foreground scores.

During training-time matching, category information and spatial location jointly determine candidate–target supervision assignments. Grounding DINO (Liu et al., 2024) and Open-Det (Cao et al., 2025) compute semantic matching scores by aligning decoder queries with text features, without incorporating the additional regional semantics aggregated by our namer. To this end, the **Semantics-Guided Object Detector** incorporates the enriched category semantics in z_i into training-time matching, with the semantic cost measuring candidate–category compatibility and the geometric costs accounting for spatial correspondence.

Specifically, at decoder layer ℓ, let $Z^\ell = [z_1^\ell, \ldots, z_N^\ell]^\top \in \mathbb{R}^{N \times d_s}$ denote the candidate semantic-query matrix. The frozen VLM text encoder embeds the category names of the M ground-truth objects into $T = [t_1, \ldots, t_M]^\top \in \mathbb{R}^{M \times d_s}$. We construct the semantic cost matrix from their cosine distances as

$$P^\ell = \mathbf{1}_{N \times M} - \nu(Z^\ell)\nu(T)^\top \in \mathbb{R}^{N \times M}, \tag{2}$$

where ν denotes row-wise ℓ_2 normalization and $\mathbf{1}_{N \times M}$ is an all-ones matrix. The matrix computes distances between all candidates and ground-truth categories within each image, with a smaller P_{ij}^ℓ indicating stronger compatibility between candidate i and the category of target j.

We then calibrate P^ℓ to the scale of the original detector's semantic matching cost, obtaining $C^{\mathrm{sem},\ell}$, and combine it with the original weighted box and foreground costs for Hungarian matching. This strategy is applied in all six decoder layers, each using its own candidate features and predicted boxes while retaining the original matching weights and assignment multiplicities. The matching costs and assignments are computed without gradients, and the resulting pairs are used by the existing training losses. Scale calibration, numerical safeguards, and layer-wise assignment details are provided in Appendix A.2.

3.3 SEMANTICALLY CONDITIONED MASK DECODER

The predicted box supplies location cues but does not explicitly convey the candidate's category semantics. When the box includes neighboring objects or imprecisely covers the target, identifying the intended region from the box prompt alone is difficult. To address this issue, our **Semantically Conditioned Mask Decoder** injects the semantic query z_i into a pretrained mask decoder from the SAM family (Kirillov et al., 2023; Ravi et al., 2025), allowing object semantics, spatial prompts, and local visual evidence to jointly guide mask prediction. The semantic query used for category generation also conditions mask decoding, allowing both tasks to share the same candidate-specific semantic representation without re-encoding generated category names.

Specifically, let $\mathbf{U} \in \mathbb{R}^{K \times d_m}$ contain the K pretrained mask tokens as rows, where d_m is the decoder token dimension. A lightweight MLP adapter $A : \mathbb{R}^{d_s} \to \mathbb{R}^{d_m}$ maps z_i into this token space, and $e_0 \in \mathbb{R}^K$ is a one-hot vector selecting the first mask token. Semantic injection and conditional decoding are expressed jointly as

$$\begin{aligned} \widetilde{\mathbf{U}}_i &= \mathbf{U} + \alpha e_0 A(z_i)^\top, \\ \{(L_{ik}, \widehat{v}_{ik})\}_{k=0}^{K-1} &= D_{\text{mask}}\Big(F^{\text{seg}}, E_b(b_i), \widetilde{\mathbf{U}}_i\Big), \end{aligned} \tag{3}$$

where α controls the semantic injection strength, E_b encodes the box prompt, and D_{mask} denotes the mask decoder. The adapted semantic vector is added to the first mask token before token–image interaction; the other mask tokens receive no direct semantic update. The decoder retains its original quality and other output tokens and outputs mask logits L_{ik} and corresponding quality scores $\widehat{v}_{ik}$. Adapter and readout details are provided in Appendix A.3; training objectives and quality-based mask selection are described in Section 3.4.

3.4 TRAINING AND INFERENCE

We train UniOSE in two stages. The first stage jointly trains the detector and object namer using box and category-name annotations, while freezing the shared image encoder and VLM image and text encoders. The second stage freezes the detection and naming model and adapts segmentation using mask annotations, training the semantic adapter and copied mask and quality readouts, followed by selective low-rank decoder fine-tuning (Hu et al., 2022).

Detection and naming training. The detector and object namer are jointly optimized with box-regression, foreground-classification, semantic-alignment, distillation, and category-name generation objectives. The semantics-guided matching in Section 3.2 determines candidate–target supervision assignments. Direct semantic-alignment and naming supervision is applied at the final two detector decoder layers, while semantics-guided matching is used at all six layers.

Segmentation training. The segmentation objective $\mathcal{L}_{\text{seg}}$ combines Dice and focal losses (Lin et al., 2017) for mask prediction with a regression loss supervising the predicted quality scores. This objective is used for segmentation adaptation and subsequent low-rank fine-tuning. The detection and naming model, shared image encoder, and original mask and quality readouts remain frozen throughout segmentation training. Mask supervision and readout details are provided in Appendix A.3.

Inference. Given only an image, all dynamically selected candidates undergo semantic-query construction and name generation before candidate suppression. The retained boxes and semantic queries condition mask decoding; we select the mask with the highest predicted quality and restore it to the original image coordinates. The output $\mathcal{O} = \{(b_i, y_i, m_i) : i \in \mathcal{S}\}$ preserves object correspondence across boxes, names, and masks. Ground-truth text and Hungarian matching are used only during training.

Table 1: Open-ended detection on LVIS minival. Vocabulary-free indicates inference without a supplied category vocabulary (✓). — denotes unreported values.

Method	Epochs	Vocabulary-free	Backbone	Pre-Train Data	LVIS minival			
					AP_r	AP_c	AP_f	AP
Grounding DINO	—	×	Swin-T	O365, GoldG, CAP4M	18.1	23.3	32.7	27.4
YOLO-World-L	—	×	YOLOv8-L	O365, GoldG, CC3M-Lite	27.6	34.1	38.0	35.4
LLMDet	—	×	Swin-T	GroundingCap-1M	37.3	39.5	50.7	44.7
GenerateU	149	✓	Swin-T	VG	17.4	22.4	29.6	25.4
GenerateU	—	✓	Swin-T	VG + GRIT5M	20.0	24.9	29.8	26.8
Open-Det	31	✓	Swin-T	VG	21.0	24.8	30.1	27.0
Open-Det	50	✓	Swin-T	VG	21.9	25.1	30.4	27.4
VL-SAM-V2	—	✓	Swin-T	GroundingCap-1M	**29.8**	—	—	29.5
UniOSE (ours)	5	✓	SAM2-T	VG	19.5	28.6	31.6	$29.2_{\uparrow 1.8}$
UniOSE (ours)	15	✓	SAM2-T	VG	28.5	**31.5**	**33.9**	$\mathbf{32.4}_{\uparrow 5.0}$
GenerateU	—	✓	Swin-L	VG + GRIT5M	22.3	25.2	31.4	27.9
Open-Det	31	✓	Swin-L	VG	**31.2**	32.1	34.3	33.1
VL-SAM-V2	—	✓	Swin-L	GroundingCap-1M	30.5	—	—	31.8
UniOSE (ours)	15	✓	SAM2-L	VG	26.5	**36.5**	**38.6**	$\mathbf{36.7}_{\uparrow 3.6}$

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. Detection and naming use Visual Genome (VG) (Krishna et al., 2017) for task training, while segmentation adaptation uses 50,000 images with thing and stuff annotations from COCO Panoptic (Lin et al., 2014; Kirillov et al., 2019). We match the trained detector's predictions with ground-truth annotations to construct object-level paired training samples. Detection is evaluated on LVIS minival (Gupta et al., 2019), and zero-shot semantic segmentation is evaluated on ADE20K with 847 (A-847) and 150 (A-150) categories (Zhou et al., 2017), PASCAL Context with 459 (PC-459) and 59 (PC-59) categories (Mottaghi et al., 2014), and the PASCAL VOC 20-category setting (VOC-20) (Everingham et al., 2015). Open-ended methods generate category names from the image before mapping them to benchmark labels. For detection, we report AP and category-frequency breakdowns AP_r, AP_c, and AP_f on LVIS minival. Semantic segmentation is evaluated by mean intersection-over-union (mIoU). For the diagnostic analyses in Section 4.4, we report semantic query–GT alignment scores and Joint@0.50/Joint@0.75; the Joint metrics are defined in Appendix B.1.

Implementation Details. UniOSE uses a SAM 2 backbone (Ravi et al., 2025), an EVA02-CLIP vision–language encoder (Fang et al., 2024; Sun et al., 2023), and a FLAN-T5-Base name generator (Raffel et al., 2020; Chung et al., 2024). Detection uses the Tiny and Large SAM 2 variants, while all segmentation experiments use the Tiny variant. Experiments are conducted on four NVIDIA RTX 6000 Ada GPUs. We use AdamW (Loshchilov & Hutter, 2019) with weight decay 0.05 and an effective global batch size of 8 images for detection and naming, and weight decay 10^{-4} with an effective global batch size of 64 masks for segmentation training. The base learning rate for detection and naming is 10^{-4}, with separate rates for the name generator and selected detection modules. Segmentation uses a learning rate of 10^{-4} for the semantic adapter and 10^{-5} for the mask and quality readouts and subsequent LoRA updates. Sensitivity to the evaluation confidence-fusion weight is examined in Appendix B.2.

4.2 Comparison with State-of-the-Art Methods

Open-ended Detection. Table 1 compares UniOSE with open-ended detectors on LVIS minival, with vocabulary-prompted OVD methods listed separately for reference. With VG task training, UniOSE with SAM2-T achieves 29.2 AP after only five epochs, exceeding GenerateU's 25.4 AP after 149 epochs (Lin et al., 2024a) and Swin-T Open-Det's 27.4 AP after 50 epochs (Cao et al.,

Table 2: Zero-shot open-ended semantic segmentation results in mIoU. ESC-Net (Lee et al., 2025) is included as a vocabulary-prompted reference. Parameters refer to the listed naming component, not the full model; GPT-2 (Radford et al., 2019) and FLAN-T5 counts exclude visual encoders, while BLIP and CLIP counts include both visual and text components. DB denotes database retrieval; — indicates an unavailable or inapplicable value. UniOSE uses SAM2-T.

Method	Vocabulary-free	Naming model	Params (M)	mIoU				
				A-847	PC-459	A-150	PC-59	VOC-20
ESC-Net	×	—	—	18.1	27.0	41.8	65.6	98.3
ZeroSeg	✓	GPT-2 Medium	355	—	11.4	—	—	20.1
TAG	✓	CLIP-L/14 + DB	428	—	—	6.6	20.2	56.9
Chicken-and-egg	✓	RAM (Swin-L)	—	6.7	8.0	18.8	27.8	81.8
DenseCaSED	✓	CLIP-L/14 + DB	428	—	—	9.8	15.9	24.2
AutoSeg (LAVE)	✓	BLIP-L	470	6.0	11.7	—	—	87.1
AutoSeg (Manual)	✓	BLIP-L	470	6.2	12.8	—	—	**88.2**
UniOSE (ours)	✓	FLAN-T5-Base	248	**11.1**	**13.4**	**31.8**	**46.0**	78.1

Table 3: Effects of the VLM branch and semantic-query matching on LVIS minival after 5 epochs.

VLM branch	Semantic matching	AP_r	AP_c	AP_f	AP
×	×	8.2	10.3	23.3	16.4
×	✓	5.8	11.1	23.1	16.5
✓	×	**20.3**	26.3	**31.7**	28.4
✓	✓	19.5	**28.6**	31.6	**29.2**

Table 3 notes. VLM branch: × = Base Query, ✓ = base query + VLM. Matching: × = decoder query–GT, ✓ = fused semantic query.

Table 4: Effect of semantic-query conditioning on segmentation.

Conditioning	A-847	PC-459	VOC-20
bbox-only	10.8	13.0	77.0
Random query	10.8	12.8	74.6
Empty query	10.5	12.3	62.7
Ours	**11.1**	**13.4**	**78.1**

2025) by 3.8 and 1.8 points, respectively. After 15 epochs, UniOSE reaches 32.4 AP, increasing these margins to 7.0 and 5.0 points. The improvement also holds for the larger configuration, where UniOSE with SAM2-L achieves 36.7 AP after 15 epochs, surpassing Swin-L Open-Det's 33.1 AP after 31 epochs by 3.6 points. Both configurations thus achieve higher AP with fewer task-training epochs. VL-SAM-V2 (Lin & Wang, 2025) provides a further comparison of detection performance and naming-model scale. Our Tiny and Large configurations obtain 32.4 and 36.7 AP, while VL-SAM-V2 reports scores of 29.5 and 31.8 for its corresponding configurations. UniOSE uses a FLAN-T5-Base name generator with approximately 248M parameters, whereas VL-SAM-V2 incorporates InternVL2.5-8B alongside its detection components. UniOSE therefore obtains higher reported scores with a substantially smaller name-generation component.

Open-ended Semantic Segmentation. Table 2 evaluates the transfer of UniOSE's object representations to zero-shot semantic segmentation across five benchmark settings. All segmentation experiments use SAM2-T and 50,000 COCO Panoptic images for segmentation adaptation. Under this setting, UniOSE outperforms the strongest baseline, Chicken-and-egg with RAM tagging (Reichard et al., 2025), by 4.4, 13.0, and 18.2 mIoU points on A-847, A-150, and PC-59, respectively; on PC-459, it exceeds AutoSeg (Manual) (Ülger et al., 2025) by 0.6 points. These results demonstrate effective transfer of the detection-side object representations to semantic segmentation using SAM2-T and 50K-image mask adaptation.

4.3 Ablation Studies

Semantic Representation and Matching. Table 3 evaluates VLM region features and semantic-query matching after five training epochs, retaining the object namer in all configurations. The VLM branch supplements detector-derived object representations with regional semantics; matching

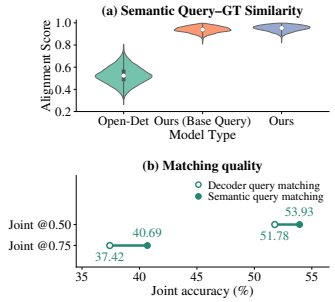


Figure 4: Effect of semantic queries on alignment and matching. See Appendix B.1 for source statistics and definitions.

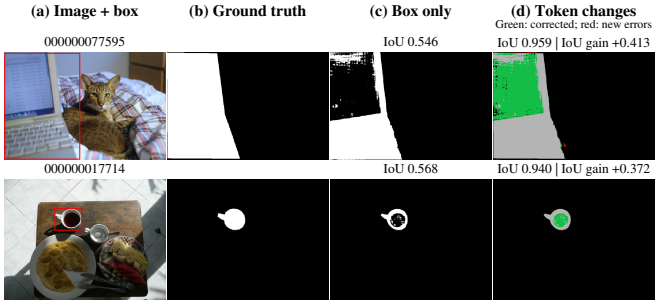


Figure 5: Semantic-query conditioning with fixed images and boxes. Panel (d) shows changes from the box-only mask in (c): green marks corrected pixels; red marks new errors. Its IoU and gain refer to the conditioned mask.

uses either the original decoder query or the fused semantic query to compute the semantic cost. Adding VLM features improves AP from 16.4 to 28.4 with decoder-query matching and from 16.5 to 29.2 with semantic-query matching. Switching to semantic-query matching yields gains of 0.1 AP without VLM features and 0.8 AP with them, suggesting that richer object semantics make matching more effective. The full model reaches 29.2 AP, with common-category AP increasing from 26.3 to 28.6, although rare-category AP decreases from 20.3 to 19.5.

Semantic Conditioning for Segmentation. Table 4 compares *bbox-only*, *Random query*, *Empty query*, and *Ours*. Adding the object's semantic query z_i to the box prompt improves mIoU over bbox-only by 0.3, 0.4, and 1.1 points on A-847, PC-459, and VOC-20, respectively. Random queries do not outperform bbox-only on any benchmark, while empty queries further reduce performance, reaching 62.7 mIoU on VOC-20. Ours outperforms both query controls across all three benchmarks, supporting the value of object-specific semantics beyond introducing an additional query.

4.4 Analysis

Semantic Alignment and Matching Quality. Figure 4(a) compares query–GT alignment on jointly localized targets from the same 500 images. Mean scores are 0.53 for Open-Det, 0.94 for our model without the VLM branch, and 0.95 for the full model (Table 6). Figure 4(b) compares two full models trained for 50K iterations with decoder-query or semantic-query matching. Joint@τ requires correct naming and localization at IoU $\geq \tau$ (Appendix B.1). Semantic-query matching improves Joint@0.50 from 51.78% to 53.93% and Joint@0.75 from 37.42% to 40.69%, gains of 2.16 and 3.27 percentage points (Table 7). Both source tables and metric definitions are provided in Appendix B.1.

Qualitative Segmentation Analysis. Figure 5 shows two selected examples with the image and predicted box held fixed. Semantic-query conditioning recovers missing interior regions in the laptop and cup, increasing mask IoU from 0.55 to 0.96 and from 0.57 to 0.94, respectively. These cases illustrate mask refinement beyond box prompts; additional qualitative visualizations are provided in Appendix B.3.

5 Conclusion

We presented UniOSE for open-ended object detection and segmentation without a predefined category vocabulary. UniOSE enriches detector-derived object queries with regional VLM semantics to support category-name generation, and uses the same semantic queries to guide training-time detection matching and mask decoding. This design connects category generation with spatial prediction through a shared object representation. Experiments show a 5.0 AP improvement over Swin-T Open-Det on LVIS minival using SAM2-T, together with effective zero-shot transfer to semantic segmentation across five benchmark settings. Component ablations support the contributions of regional semantic features and semantic-query guidance to detection and segmentation.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting, translation, language editing, and organization. The authors take full responsibility for the final manuscript, including all AI-assisted text, scientific claims, and reported results.

REPRODUCIBILITY STATEMENT

The proposed method and training procedure are described in Section 3, and the experimental setup and evaluation settings are outlined in Section 4.1. Appendix A provides additional implementation details for semantic matching and mask adaptation, while Section 4.3 presents the component ablation design. We will make the code and model checkpoints publicly available upon acceptance to facilitate reproduction and further research.

REFERENCES

Guiping Cao, Tao Wang, Wenjian Huang, Xiangyuan Lan, Jianguo Zhang, and Dongmei Jiang. Open-Det: An Efficient Learning Framework for Open-Ended Detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6654–6674. PMLR, 2025.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Computer Vision – ECCV 2020*, pp. 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-Attention Mask Transformer for Universal Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299, 2022.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. CAT-Seg: Cost Aggregation for Open-Vocabulary Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4113–4123, 2024.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Achal Dave, Piotr Dollár, Deva Ramanan, Alexander Kirillov, and Ross Girshick. Evaluating Large-Vocabulary Object Detectors: The Devil is in the Details. *arXiv preprint arXiv:2102.01066*, 2021.

Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-Vocabulary Universal Image Segmentation with MaskCLIP. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8090–8102. PMLR, 2023.

Debidatta Dwibedi, Vidhi Jain, Jonathan Tompson, Andrew Zisserman, and Yusuf Aytar. Flex-Cap: Describe Anything in Images in Controllable Detail. In *Advances in Neural Information Processing Systems*, volume 37, pp. 111172–111198. Curran Associates, Inc., 2024. doi: 10.52202/079017-3530.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. doi: 10.1007/s11263-014-0733-5.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. doi: 10.1016/j.imavis.2024.105171.

Cristina González, Santiago Rodríguez, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, and Pablo Arbeláez. Benchmarking Open-ended Segmentation. In *International Conference on Learning Representations*, volume 2026, pp. 94499–94517, 2026.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *International Conference on Learning Representations*, 2022.

Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. RegionGPT: Towards Region Understanding Vision Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13796–13806, 2024.

Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5356–5364, 2019.

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9404–9413, 2019.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. doi: 10.1007/s11263-016-0981-7.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning Segmentation via Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, 2024.

Minhyeok Lee, Suhwan Cho, Jungho Lee, Sunghun Yang, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Effective SAM Combination for Open-Vocabulary Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26081–26090, 2025.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven Semantic Segmentation. In *International Conference on Learning Representations*, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded Language-Image Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10965–10975, 2022c.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-Vocabulary Semantic Segmentation With Mask-Adapted CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7070, 2023.

Chuang Lin, Yi Jiang, Lizhen Qu, Zehuan Yuan, and Jianfei Cai. Generative Region-Language Pretraining for Open-Ended Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13958–13968, 2024a.

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 *Computer Vision – ECCV 2014*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

Zhiwei Lin and Yongtao Wang. VL-SAM-V2: Open-World Object Detection with General and Specific Query Fusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4485.

Zhiwei Lin, Yongtao Wang, and Zhi Tang. Training-Free Open-Ended Object Detection and Segmentation via Attention as Prompts. In *Advances in Neural Information Processing Systems*, volume 37, pp. 69588–69606. Curran Associates, Inc., 2024b. doi: 10.52202/079017-2223.

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 *Computer Vision – ECCV 2024*, pp. 38–55. Springer, 2024. doi: 10.1007/978-3-031-72970-6_3.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple Open-Vocabulary Object Detection with Vision Transformers. In *Computer Vision – ECCV 2022*, pp. 728–755. Springer, 2022. doi: 10.1007/978-3-031-20080-9_42.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling Open-Vocabulary Object Detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 72983–73007. Curran Associates, Inc., 2023. doi: 10.52202/075280-3191.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 891–898, 2014.

Ting Pan, Lulu Tang, Xinlong Wang, and Shiguang Shan. Tokenize Anything via Prompting. In *Computer Vision – ECCV 2024*, pp. 330–348. Springer, 2024. doi: 10.1007/978-3-031-72970-6_19.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI, 2019.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Klara Reichard, Giulia Rizzoli, Stefano Gasperini, Lukas Hoyer, Pietro Zanuttigh, Nassir Navab, and Federico Tombari. From Open-Vocabulary to Vocabulary-Free Semantic Segmentation. *Pattern Recognition Letters*, 198:14–21, 2025. doi: 10.1016/j.patrec.2025.08.025.

Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, Xingyu Chen, Zhuheng Song, Yuhong Zhang, Hongjie Huang, Han Gao, Shilong Liu, Hao Zhang, Feng Li, Kent Yu, and Lei Zhang. DINO-X: A Unified Vision Model for Open-World Object Detection and Understanding. *arXiv preprint arXiv:2411.14347*, 2024.

Pitchaporn Rewatbowornwong, Nattanat Chatthee, Ekapol Chuangsuwanich, and Supasorn Suwajanakorn. Zero-guidance Segmentation Using Zero Segment Labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1162–1172, 2023.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, 2019.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv preprint arXiv:2303.15389*, 2023.

Osman Ülger, Maksymilian Kulicki, Yuki Asano, and Martin R. Oswald. Auto-Vocabulary Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24266–24275, 2025.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side Adapter Network for Open-Vocabulary Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2945–2954, 2023.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and Ground Anything Anywhere at Any Granularity. In *International Conference on Learning Representations*, volume 2024, pp. 57153–57180, 2024.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions Die Hard: Open-Vocabulary Segmentation with Single Frozen Convolutional CLIP. In *Advances in Neural Information Processing Systems*, volume 36, pp. 32215–32234. Curran Associates, Inc., 2023. doi: 10.52202/075280-1399.

Qihang Yu, Xiaohui Shen, and Liang-Chieh Chen. Towards Open-Ended Visual Recognition with Large Language Models. In *Computer Vision – ECCV 2024*, pp. 359–376. Springer, 2024. doi: 10.1007/978-3-031-72630-9_21.

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel Understanding with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28202–28211, 2024.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-Vocabulary DETR with Conditional Matching. In *Computer Vision – ECCV 2022*, 2022.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-Vocabulary Object Detection Using Captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14393–14402, 2021.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In *International Conference on Learning Representations*, 2023a.

Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A Simple Framework for Open-Vocabulary Segmentation and Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1020–1031, 2023b.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-Based Language-Image Pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16793–16803, 2022.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing Through ADE20K Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, 2017.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting Twenty-Thousand Classes Using Image-Level Supervision. In *Computer Vision – ECCV 2022*, pp. 350–368. Springer, 2022. doi: 10.1007/978-3-031-20077-9_21.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*, 2021.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized Decoding for Pixel, Image, and Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15116–15127, 2023.

A Additional Method Details

A.1 Diagnostic Protocol and Results for Figure 1

Experimental design. Figure 1(c) compares the released Open-Det Swin-T checkpoint (Cao et al., 2025) with a fully fine-tuned SAM2-T backbone variant, both trained for 300K iterations. The latter uses SAM 2 Tiny initialization (Ravi et al., 2025) without the added FPN; it is the baseline denoted by "SAM" in the figure. We analyze their saved predictions on the same fixed subset of 1,000 LVIS minival images (Gupta et al., 2019) (seed 20260716), containing 11,258 GT instances and evaluated in ten disjoint 100-image shards. Both runs retain the complete dynamic-query path: selection uses a confidence threshold of 0.05, at least 300 and at most 900 queries, and every selected query undergoes visual–language mapping and T5 generation. Class-agnostic NMS at IoU 0.9 is applied after generation when the query count exceeds 300, followed by the retained Top-15 category-mapping outputs. For diagnosis, records with identical box coordinates after rounding to two decimal places are grouped, and all resulting unique boxes are retained without an additional top-300 filter.

Table 5: Source results for Figure 1(c), using two 300K-iteration checkpoints on the fixed LVIS 1,000-image subset. All values are percentages, including mean IoU multiplied by 100. Top-1, Top-15, and description hit are conditional on localization at IoU ≥ 0.5. Joint$_{15}$ is the any-box, Top-15 recall defined below. Bold marks the higher value in each column.

Model	Mean IoU	Loc. @0.50	Loc. @0.75	Category Top-1	Category Top-15	Description hit	Joint$_{15}$ @0.50	Joint$_{15}$ @0.75
Open-Det Swin-T	71.33	80.09	61.51	**62.11**	**88.85**	**68.42**	**73.38**	**56.08**
SAM2-T (full FT)	**71.78**	**81.04**	**62.27**	58.77	86.06	64.13	72.04	55.17

Metric definitions. For each GT instance, we select its highest-IoU predicted box independently, without one-to-one assignment. *Mean IoU* averages these maximum IoUs over all GT instances, assigning zero when no box is available; *Loc.*@τ is the fraction whose maximum IoU reaches τ. *Category Top-1* checks whether the highest-scoring mapped category on the selected box equals the GT category; *Top-15* checks whether its retained category set contains the GT category. Both are averaged only over GT instances with maximum IoU ≥ 0.5, separately for each model. *Description hit* uses the same localized subset and checks whether any generated description associated with the selected box contains a GT category name or synonym as a whole-word sequence after lowercasing and normalizing non-alphanumeric characters to spaces. *Joint*$_{15}$@τ uses all GT instances as its denominator: a GT succeeds if *any* retained box has IoU $\geq \tau$ and contains its category in the Top-15 mapped set. It is distinct from the assigned-query, Top-1 Joint@τ metric in Appendix B.1.

Connection to Figure 1(c). The figure plots Loc.@0.75 (61.51/62.27) and conditional Category Top-1 (62.11/58.77) for Swin-T/SAM2-T, rounded to one decimal place: 61.5/62.3 and 62.1/58.8, respectively. Thus, SAM2-T improves localization recall by 0.76 percentage points while reducing conditional naming accuracy by 3.34 points in this comparison.

A.2 Semantic Matching Details

For the cosine distance P^ℓ in Equation 2, we use FP32 and row-wise normalization $\nu(x) = x/\max(\|x\|_2, 10^{-8})$. Here, N_ℓ denotes the candidate count at layer ℓ, written as N in Section 3.2. Let $D^\ell \in \mathbb{R}^{N_\ell \times M}$ be the original detector's unweighted semantic cost, computed from decoder alignment logits using sigmoid and the positive-minus-negative focal cost ($\gamma = 2$, logarithmic stabilizer 10^{-8}, and $\alpha_\ell = 2.0/0.25$ for the first four/final two layers).

We center each GT column over candidates and calibrate P^ℓ to the root mean square (RMS) of D^ℓ:

$$\begin{aligned} \mu_j^\ell(X) &= \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} X_{ij}, \qquad X_{ij}^\circ = X_{ij} - \mu_j^\ell(X), \\ r_\ell &= \frac{\mathrm{RMS}((D^\ell)^\circ)}{\max(\mathrm{RMS}((P^\ell)^\circ), \epsilon_c)}, \\ C_{ij}^{\mathrm{sem},\ell} &= \mu_j^\ell(D^\ell) + r_\ell (P^\ell)_{ij}^\circ. \end{aligned} \tag{4}$$

RMS uses all entries of the image-local matrix, with one scale per image and layer and $\epsilon_c = 10^{-6}$. We revert to D^ℓ if either centered RMS is at most ϵ_c, either input contains non-finite values, r_ℓ is non-finite or exceeds 100, or expanded target slots outnumber candidates; the scale is not clipped.

The total matching cost and assignment are

$$\begin{aligned} C_{ij}^\ell &= \lambda_{\mathrm{sem}}^\ell C_{ij}^{\mathrm{sem},\ell} + \lambda_{\mathrm{box}}^\ell \|b_i^\ell - b_j^\star\|_1 \\ &\quad - \lambda_{\mathrm{giou}}^\ell \,\mathrm{GIoU}(b_i^\ell, b_j^\star) + C_i^{\mathrm{obj},\ell}, \\ \mathcal{A}_\ell^\star &= \arg\min_{\mathcal{A}\in\Pi_\ell} \sum_{(i,j)\in\mathcal{A}} C_{ij}^\ell, \end{aligned} \tag{5}$$

where $b_j^\star$ is the GT box, GIoU follows Rezatofighi et al. (2019), $C_i^{\mathrm{obj},\ell}$ is the inherited foreground cost, and Π_ℓ contains maximum-cardinality one-to-one assignments to target slots. We retain $\lambda_{\mathrm{sem}}^\ell = 0.2/2.0$ for the first four/final two layers, $\lambda_{\mathrm{box}}^\ell = 5$, $\lambda_{\mathrm{giou}}^\ell = 2$, and unit foreground weight.

Matching uses only valid within-image text columns; images without GT are skipped. The first four layers duplicate each GT into two slots: calibration precedes duplication, costs follow the box order, and the fallback guard counts both slots. The final two layers use one-to-one matching. Readers and fusion are shared across layers; early layers reuse the final two layers' shared semantic mapping head. Calibration and assignment require no gradients or new parameters; encoder-proposal matching, denoising assignments, and training losses remain unchanged.

A.3 Mask Adapter and Decoder Details

We map z_i into the mask decoder's token space through a lightweight adapter:

$$\begin{aligned} a_i^{\mathrm{sem}} &= A(z_i) = \mathrm{LN}(W_{a,2}\,\mathrm{GELU}(W_{a,1}\,\mathrm{LN}(z_i) + c_1) + c_2)\,, \\ \widetilde{u}_{i,0} &= u_0 + \alpha_{\mathrm{sem}} a_i^{\mathrm{sem}}, \end{aligned} \tag{6}$$

where $W_{a,1} \in \mathbb{R}^{d_a \times d_s}$ and $W_{a,2} \in \mathbb{R}^{d_m \times d_a}$, $c_1 \in \mathbb{R}^{d_a}$ and $c_2 \in \mathbb{R}^{d_m}$ are biases, d_a is the hidden dimension, d_m is the decoder token dimension, and LN denotes layer normalization. The adapted vector is added to the first mask token u_0 before the bidirectional attention blocks. The coefficient α_{sem} (denoted by α in Section 3.3) increases from zero to one during adaptation. Other mask tokens and the quality token retain their pretrained initialization and positions.

As in Section 3.3, $\widetilde{\mathbf{U}}_i \in \mathbb{R}^{K \times d_m}$ contains only the mask tokens, with first row $\widetilde{u}_{i,0}^\top$; the decoder internally adds the quality token and any other output tokens. The box-prompt encoder is denoted by E_b. The decoder produces candidate masks and quality scores as

$$\{(L_{ik}, \widehat{v}_{ik})\}_{k=0}^{K-1} = D_{\mathrm{mask}}\Big(F^{\mathrm{seg}}, E_b(b_i), \widetilde{\mathbf{U}}_i\Big)\,, \tag{7}$$

where L_{ik} denotes mask logits and $\widehat{v}_{ik}$ is their predicted quality. The box is transformed to the prompt encoder's coordinate system before embedding. Through token–image interaction, the injected semantics can influence the mask representation alongside the image features and box prompt.

We initialize trainable copies of the first mask hypernetwork and its associated quality readout from the pretrained decoder; the remaining masks use the original readouts. Supervision and quality-based mask selection are described in Section 3.4. Because the semantic condition comes from the candidate's pre-generation representation, name generation and mask decoding retain a common object reference without requiring a text prompt from the user.

Inherited mask-decoder supervision. The base objective $\mathcal{L}_{\text{seg}}$ retains the original mask-decoder supervision. For each region, the mask candidate with the lowest weighted Dice-plus-focal loss is selected for training; denote its index by $k_i^\star$. The selected mask receives mask supervision and a squared-error quality objective against its detached IoU with the ground-truth mask. This ground-truth-based candidate selection is used only during training.

B Additional Experimental Details

B.1 Diagnostic Protocol and Results for Figure 4

Figure 4 uses a fixed set of 500 LVIS minival images with 5,286 non-crowd GT instances (seed 20260924; five 100-image shards). Tables 6 and 7 give the source statistics for panels (a) and (b), respectively. Both diagnostics retain all dynamically selected queries before NMS; complete T5 generation is verified against the native inference outputs.

Semantic alignment: models and target selection. Panel (a) compares the released Open-Det Swin-T 300K checkpoint (Cao et al., 2025) with two SAM2-T models trained for 50K iterations using decoder-query matching. Ours (Base Query) fuses decoder-query, encoder-memory region, and detection-feature region representations; Ours additionally includes the VLM region branch. Both retain the name generator. For each model and image, we find a one-to-one query–GT assignment by maximizing total box IoU, independently of semantic scores. We then retain the intersection $\mathcal{G}_\cap$ of GT instances assigned a box with IoU ≥ 0.5 by all three models: $|\mathcal{G}_\cap| = 3,966$ across the same 500 images. This shared subset is used only for panel (a).

Alignment score and distribution statistics. Let $z_j^{(m)}$ be model m's fused visual query supplied to the name generator for the query assigned to GT j, and $t_{c_j}^{(m)}$ its native text embedding of the GT category. We encode the canonical category name, replacing underscores with spaces and using the template "a {category name}". With ν denoting ℓ_2 normalization, the plotted score and reported mean are

$$a_j^{(m)} = \frac{1 + \nu(z_j^{(m)})^\top \nu(t_{c_j}^{(m)})}{2}, \qquad \bar{a}^{(m)} = \frac{1}{|\mathcal{G}_\cap|} \sum_{j \in \mathcal{G}_\cap} a_j^{(m)}. \tag{8}$$

Scores lie in $[0, 1]$ and measure positive-category alignment, not naming accuracy. Table 6 reports the mean, sample standard deviation (SD, denominator $n - 1$), median, and 25th/75th percentiles (Q1/Q3, linearly interpolated). In the violin plot, white dots mark medians and thick bars span Q1–Q3; thin bars extend to the 1.5-IQR limits, clipped to the observed minimum and maximum. Open-Det uses its native OpenAI CLIP text space, whereas both UniOSE variants use EVA02. Scores are computed within each model's native text space, without cross-space calibration.

Table 6: Alignment-score statistics underlying Figure 4(a), computed on the same 3,966 jointly localized GT instances. All scores use Equation 8; the means rounded to two decimals give the values in the main text.

Model	Mean	SD	Median	Q1	Q3
Open-Det	0.5264	0.0786	0.5240	0.4733	0.5776
Ours (Base Query)	0.9385	0.0278	0.9396	0.9212	0.9589
Ours	0.9503	0.0264	0.9531	0.9337	0.9702

Joint localization and category accuracy. Panel (b) compares two four-branch SAM2-T checkpoints independently trained for 50K iterations, one with decoder-query matching and the other with semantic-query matching. Each is evaluated with its corresponding matcher on the final decoder layer. The matching cost combines the category term, box L1, GIoU, and foreground cost; semantic-query matching uses the RMS calibration and fallback in Appendix A.2. The query–GT assignment is obtained offline with GT information and uses all pre-NMS queries, including queries suppressed in the final detection output. Generated names for these queries are mapped with the same native category mapper; correctness uses its Top-1 category, not membership in the Top-15 set.

Let $N_{\mathrm{GT}} = 5,286$ and $\pi(j)$ be the query assigned to GT instance j. Its predicted box and mapped Top-1 category are $b_{\pi(j)}$ and $\hat{c}_{\pi(j)}$; the GT box and category are b_j^{GT} and c_j^{GT}. We define

$$\mathrm{Joint@}\tau = \frac{100}{N_{\mathrm{GT}}} \sum_{j=1}^{N_{\mathrm{GT}}} \mathbf{1}\left[\mathrm{IoU}(b_{\pi(j)}, b_j^{\mathrm{GT}}) \geq \tau\right] \mathbf{1}\left[\hat{c}_{\pi(j)} = c_j^{\mathrm{GT}}\right]. \tag{9}$$

An unassigned GT contributes zero. Both conditions must hold for the same assigned query and GT. Unlike panel (a), the denominator includes all evaluated GT instances, rather than only the jointly localized subset. This is an assignment diagnostic, distinct from official detection AP and the any-box Top-15 recall in Appendix A.1.

Table 7: Counts and joint accuracies underlying Figure 4(b). Both rows use all 5,286 GT instances from the same 500 images. Each row uses a separately trained 50K checkpoint and its native matching strategy.

Matching strategy	Correct GT IoU ≥ 0.50	Joint@0.50 (%)	Correct GT IoU ≥ 0.75	Joint@0.75 (%)
Decoder-query matching	2,737	51.78	1,978	37.42
Semantic-query matching	2,851	53.93	2,151	40.69
Difference	+114	+2.16 pp	+173	+3.27 pp

Reading panel (b). The hollow and filled points denote decoder-query and semantic-query matching, respectively; connecting segments join absolute accuracies and are not confidence intervals. For example, $100 \times 2737/5286 = 51.7783\%$ and $100 \times 2851/5286 = 53.9349\%$ yield the plotted 51.78% and 53.93%. The gain is calculated before rounding: $100 \times 114/5286 = 2.1566$ percentage points, reported as 2.16. The IoU-0.75 gain is similarly $100 \times 173/5286 = 3.2728$ points.

B.2 Sensitivity to Confidence Fusion

We examine the sensitivity of UniOSE to the confidence-fusion weight using $s = p_{\text{binary}}^{\beta} s_{\text{sim}}^{1-\beta}$, where p_{binary} is the binary-head object confidence and s_{sim} is the name-to-category similarity score. Here β is a scoring weight, distinct from the mask-injection coefficient in Section 3.3. For each model, we fix the 150K-iteration checkpoint, generated candidates, and category mapping, and vary β from 0 to 1 in increments of 0.1 on the complete LVIS minival set. As shown in Figure 6, both UniOSE-T and UniOSE-L achieve their highest AP on this grid at $\beta = 0.2$, reaching 32.41 and 36.66, respectively. Compared with equal weighting ($\beta = 0.5$), this setting improves AP by 4.23 and 5.39 points. Using either score alone yields lower performance, indicating that object confidence and name similarity provide complementary ranking information. These results support our choice of $\beta = 0.2$, while showing that performance is sensitive to their relative weighting.

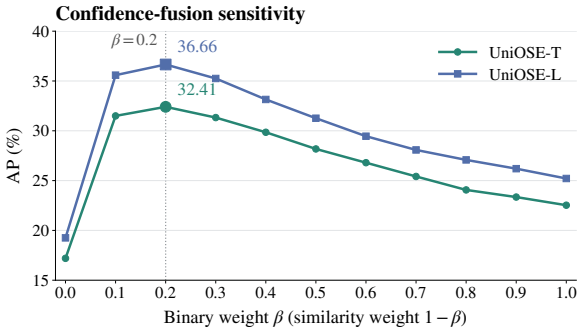


Figure 6: **Sensitivity to confidence fusion.** AP (Dave et al., 2021) on the complete LVIS minival set (4,809 images) as a function of the binary-head weight β, with the similarity weight set to $1 - \beta$. Each curve uses a fixed 150K-iteration checkpoint and identical candidates across scoring settings. Both models attain their highest observed AP at $\beta = 0.2$.

B.3 Qualitative Results

Figure 7 presents selected outputs of the final SAM2-T model on ADE-847, PC-459, and VOC-20, combining predicted boxes, generated names, and semantic segmentation maps. The examples cover indoor objects, animals, people, and vehicles.

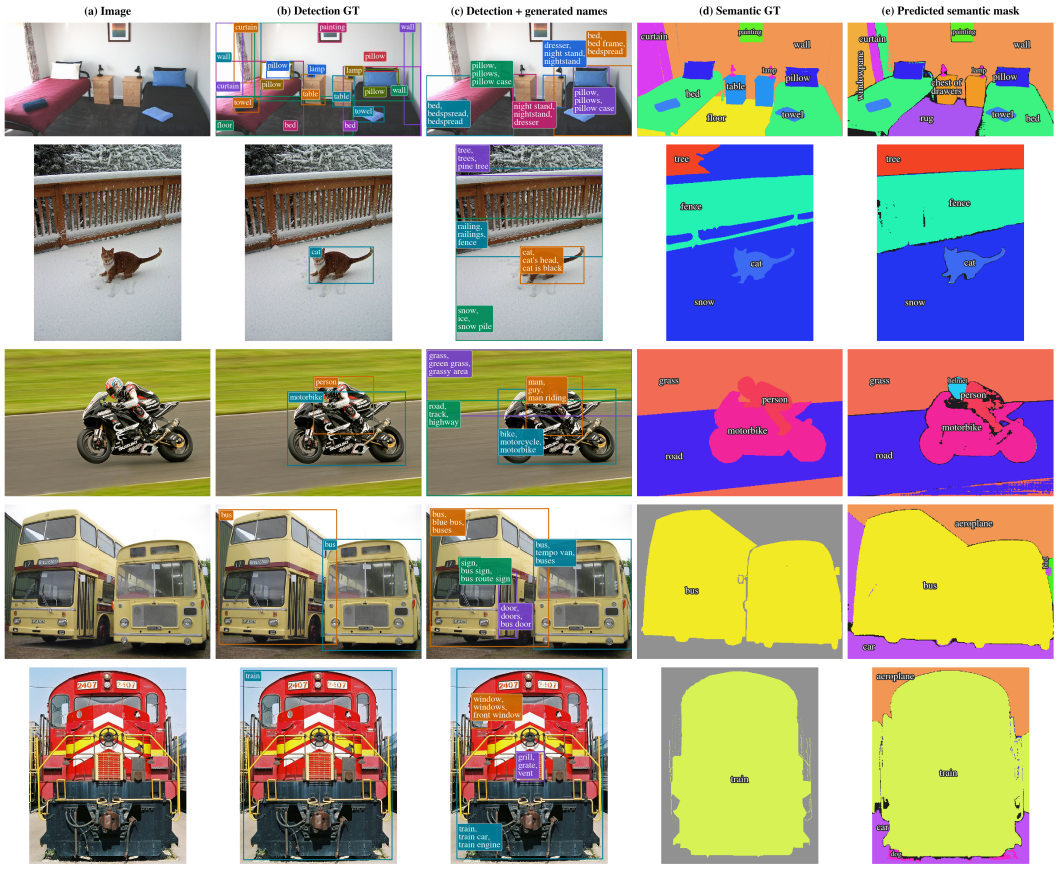


Figure 7: **Additional qualitative results of open-ended detection and semantic segmentation.** From top to bottom, examples show a twin-bed room from ADE-847, a cat in snow and a motorcycle rider from PC-459, and buses and a train from VOC-20. Columns show input images, ground-truth boxes, predicted boxes with three generated name candidates per box, ground-truth semantic maps, and predicted semantic maps. Within each example, the same semantic class uses consistent colors across the ground-truth and predicted maps. Selected detection boxes are shown for clarity, while the complete predicted semantic maps are retained. Gray regions in the ground-truth maps are ignored during evaluation.

ADE boxes are derived from instance masks; PC and VOC boxes come from the official VOC annotations. PC detection annotations cover 20 categories, whereas its semantic maps use 459 categories. Black prediction regions indicate unassigned labels.