X-CHAT: ANY PERCEPTION VIA MULTI-MODAL DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) take a perception request as a turn of dialogue, whether it names a category, describes an object, asks a question, or points at a region. Their outputs are not unified, however: segmentation MLLMs return masks alone, and MLLMs that report boxes write them as text or decode them apart from masks, each covering part of the task spectrum. We present X-Chat, an MLLM that takes multimodal dialogue as its perception interface and covers the segmentation and detection forms of seven task families (generic, open-vocabulary, referring, reasoning, grounded conversation generation, interactive, and visual grounded), fourteen tasks in all. Its language model turns the instruction into condition embeddings that drive one perceptor decoder. Each object query of that decoder regresses a box in parallel with its mask, and a geometric consistency loss couples box and mask. A detection task therefore differs from its segmentation counterpart only in which prediction is returned, and box-only data enter joint training directly. Because the existing interactive and visual grounded benchmarks score masks only, we build and release LVIS-VGD and LVIS-INT for the detection form of these two families. With one set of weights, X-Chat reports results on all fourteen tasks and is compared task by task with X-SAM and with published specialist models.

1 INTRODUCTION

A user who asks for "every car in this image", "the person in the red coat", or "everything of the same kind as the thing I clicked" wants the same output each time: a set of objects, each as a mask or a box. Today these requests go to different models with different interfaces. An open-set detector such as Grounding DINO (Liu et al., 2024b) takes a category name or a referring expression and returns boxes, and it cannot produce a mask. SAM 3 (Carion et al., 2025) segments every instance of a short noun phrase or an exemplar box, and hands referring expressions and reasoning queries to an external multimodal large language model (MLLM). Segmentation MLLMs (Lai et al., 2024; Ren et al., 2024; Rasheed et al., 2024; Zhang et al., 2024c) follow a dialogue but answer a perception request with masks alone. Figure 1 shows the alternative this paper pursues: one model that takes the request as a turn of multimodal dialogue, whether text or visual prompt, and returns whichever output form the user asks for.

Segmentation MLLMs come closest to a single interface. LISA (Lai et al., 2024) decodes a mask from the hidden state of a special token, PixelLM (Ren et al., 2024) and GLaMM (Rasheed et al., 2024) extend this to several masks per answer, and PSALM (Zhang et al., 2024c) adds category and visual-prompt conditions. X-SAM (Wang et al., 2026a) unifies seven segmentation task families in one MLLM: generic, open-vocabulary, referring, reasoning, grounded conversation generation (GCG), interactive, and visual grounded (VGD) segmentation, where a point, scribble, box, or mask prompt selects every instance of the prompted category. Its output space, however, holds masks and text only: a box enters X-SAM as a visual prompt and never leaves it as a prediction, and its training data contain no box-only annotations. So neither the detection form of these tasks nor the box-only corpora that could supervise it reaches the model.

MLLMs that do return boxes take one of three routes. UFO (Tang et al., 2025) writes each box as quantized integers in its text output and retrieves masks by matching a mask token's embedding against image features, with no task decoder at all. Groma (Ma et al., 2024) lets its language model

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.

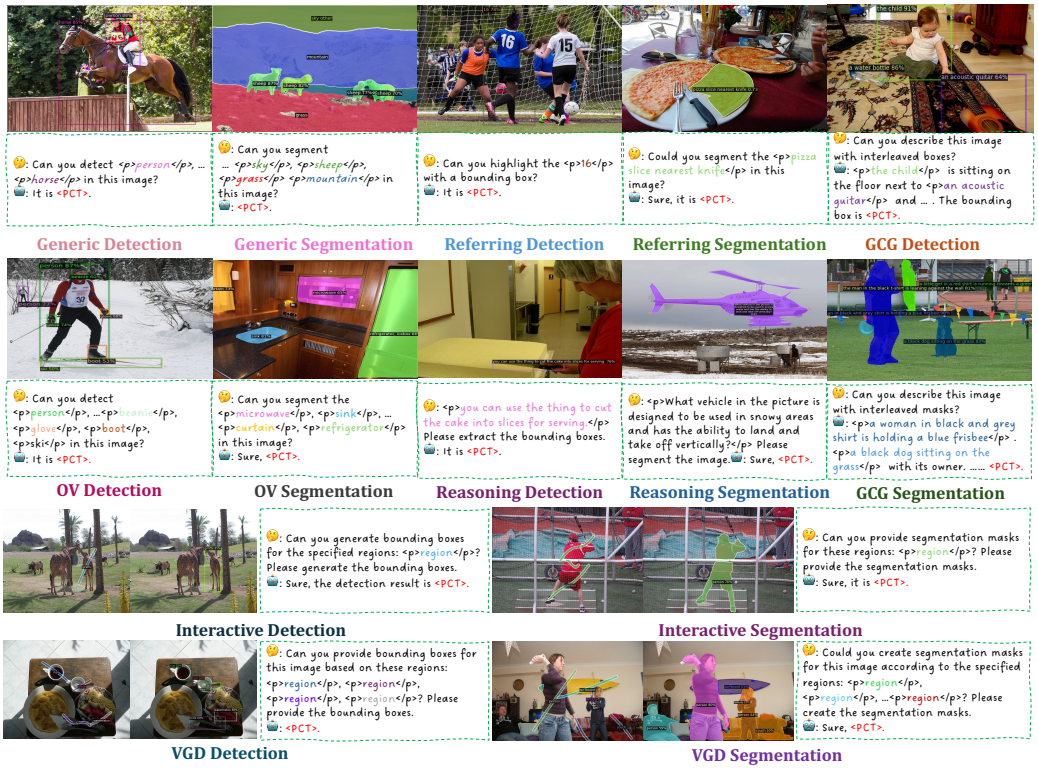

Figure 1: X-Chat answers perception requests posed as multimodal dialogue. One model covers the segmentation and detection forms of seven task families: generic, open-vocabulary (OV), referring, reasoning, grounded conversation generation (GCG), interactive, and visual grounded (VGD). The instruction names the target as category names, a referring expression, a reasoning question, a phrase of the generated answer, or a visual prompt, and states whether masks or boxes are requested. The answer marks the returned prediction with a <PCT> token.

refer to boxes from a separate region proposer through proxy tokens and produces no masks. VisionLLM v2 (Wu et al., 2024) and REF-VLM (Tai et al., 2025) route a box request and a mask request to separate task decoders, and PaDT (Su et al., 2025) decodes box and mask from one group of patch tokens through separate learnable tokens. None of them reports both forms of all seven task families; Table 1 lists what each covers.

We present X-Chat, an MLLM that takes multimodal dialogue as its perception interface and returns masks or boxes for all seven task families, fourteen tasks in all. The instruction names the target, as a category, a referring expression, a reasoning question, a phrase of the generated answer, or a visual prompt, and states the requested output form. The large language model (LLM) of X-Chat turns the instruction into condition embeddings that drive a perceptor decoder, and each object query of that decoder predicts a mask and regresses a box in parallel, as in MaskDINO (Li et al., 2023). A detection task therefore differs from its segmentation counterpart only in which prediction is returned. Samples that carry only boxes contribute their classification and box losses and skip the mask losses, which lets box-only data such as Objects365 (Shao et al., 2019) enter joint training directly. Parallel outputs alone are free to disagree, so X-Chat couples them with a geometric consistency loss between the regressed box and a differentiable box enclosing the predicted mask. In a single-seed ablation of the perceptor alone, this loss raises COCO box average precision (AP) by 1.5 over plain parallel heads at an identical parameter count. Section 4 compares it with explicit feature exchange between the mask and box predictions.

The interactive and VGD families are scored on masks by their existing benchmarks: COCO-Interactive (Zhang et al., 2024c) scores predicted masks by intersection over union, and COCO-VGD was introduced with X-SAM (Wang et al., 2026a), which outputs masks only. We therefore build LVIS-VGD and LVIS-INT on LVIS (Gupta et al., 2019) images to evaluate the detection form

Table 1: Tasks on which each MLLM reports results. ✓: the paper reports a quantitative result of the whole model on a benchmark of the task in that output form; ✗: we found no such result. Segmentation outputs masks and detection outputs boxes.

Method	*Segmentation*							*Detection*						
	Gen.	*OV*	*Ref.*	*Rea.*	*GCG*	*Int.*	*VGD*	*Gen.*	*OV*	*Ref.*	*Rea.*	*GCG*	*Int.*	*VGD*
Groma (Ma et al., 2024)	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
LocateAnything (Wang et al., 2026c)	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗
MR-MLLM (Wang et al., 2024)	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
LISA (Lai et al., 2024)	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
PixelLM (Ren et al., 2024)	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
GLaMM (Rasheed et al., 2024)	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
GROUNDHOG (Zhang et al., 2024b)	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
PSALM (Zhang et al., 2024c)	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
ARGenSeg (Wang et al., 2025)	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗
X-SAM (Wang et al., 2026a)	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
UFO (Tang et al., 2025)	✓	✗	✓	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗
SenseNova-Vision (Han et al., 2026)	✓	✗	✓	✓	✓	✓	✗	✓	✗	✓	✗	✗	✗	✗
VisionLLM v2 (Wu et al., 2024)	✓	✗	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗
REF-VLM (Tai et al., 2025)	✗	✓	✓	✗	✓	✓	✗	✗	✓	✓	✗	✓	✗	✗
PaDT (Su et al., 2025)	✗	✗	✓	✗	✗	✗	✗	✓	✗	✓	✗	✓	✗	✗
MVP-LM (Xiang et al., 2025)	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
X-Chat	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

of these two families under visual prompts, and we release both. With one set of weights, X-Chat reports results on all fourteen tasks and is compared task by task with X-SAM and with published specialist models (Table 5).

Our contributions are as follows.

- We introduce **X-Chat**, an MLLM that unifies segmentation and detection under a multi-modal dialogue interface: one model returns masks or boxes for generic, open-vocabulary, referring, reasoning, GCG, interactive, and VGD tasks, and the instruction states which output form is returned.
- We propose a **shared-query box prediction** together with benchmarks for the detection form it enables: each object query regresses a box in parallel with its mask, coupled to it by a geometric consistency loss that adds no parameters, and **LVIS-VGD** and **LVIS-INT**, which we release, evaluate the detection form of the VGD and interactive tasks, whose existing benchmarks score masks only.
- With a single set of weights, X-Chat exceeds the segmentation-only X-SAM on generic, open-vocabulary, GCG, and VGD segmentation and stays close to it on referring and reasoning segmentation, while also returning boxes for every task family.

2 Related Work

Segmentation MLLMs. LISA (Lai et al., 2024) decodes one mask from a special token's hidden state through a SAM (Kirillov et al., 2023) decoder, PixelLM (Ren et al., 2024) replaces SAM with a light pixel decoder driven by codebook tokens, and GLaMM (Rasheed et al., 2024) interleaves several segmentation tokens with a grounded description, which defines GCG. GROUNDHOG (Zhang et al., 2024b) scores externally proposed masks instead of decoding them, PSALM (Zhang et al., 2024c) conditions a mask generator on a category, a sentence, or a visual prompt, and ARGenSeg (Wang et al., 2025) generates the mask as quantized visual tokens. X-SAM (Wang et al., 2026a) unifies seven segmentation task families, including the VGD task it introduces, under one instruction format. Apart from ARGenSeg, which keeps the ability of its base MLLM to answer referring expression comprehension with a box written as text, none of them predicts a box.

MLLMs that localize with boxes. Groma (Ma et al., 2024) refers through proxy tokens to proposals from a region proposer and, as its authors state, supports no pixel-level grounding. LocateAnything (Wang et al., 2026c) decodes a box in one step as a block of quantized coordinate tokens and outputs boxes and points only. MR-MLLM (Wang et al., 2024) couples a frozen LLM with an external query-based detection head through shared queries and predicts boxes only.

Detection and unified perception MLLMs. UFO (Tang et al., 2025) writes boxes as quantized text and retrieves masks from a mask token's embedding without a task decoder; SenseNova-Vision (Han et al., 2026) also writes boxes as text but generates masks as images, under the segmentation taxonomy of X-SAM. VisionLLM v2 (Wu et al., 2024) and REF-VLM (Tai et al., 2025) route a request to a detection decoder or a separate mask decoder, and PaDT (Su et al., 2025) decodes box and mask from emitted patch tokens with separate learnable tokens over four tasks. Closest to X-Chat, MVP-LM (Xiang et al., 2025) expands an LLM token into the queries of an OpenSeeD-style decoder (Zhang et al., 2023) whose shared heads output box and mask from the same query, over generic, open-vocabulary, and referring tasks and without a reported detection AP. X-Chat covers both forms of all seven task families and couples the mask and the box of each object query with a geometric consistency loss.

Unified detection and segmentation models. MaskDINO (Li et al., 2023) adds a mask branch parallel to the box branch of DINO (Zhang et al., 2022a) and aligns the two predictions of a query through a hybrid matching cost; DI-MaskDINO (Nan et al., 2024) adds modules that rebalance the two tasks. OpenSeeD (Zhang et al., 2023) trains jointly on segmentation and box-only detection data with conditioned mask decoding, and DaTaSeg (Gu et al., 2023) supervises box-only data with a box projection loss while predicting no boxes itself. X-Decoder (Zou et al., 2023a) unifies segmentation with image-text tasks and SEEM (Zou et al., 2023b) unifies prompt types, each in one decoder that predicts masks only; GLIPv2 (Zhang et al., 2022b) casts detection as phrase grounding with per-task heads, and SAM 3 (Carion et al., 2025) segments every instance of a noun phrase or exemplar with a detector whose queries yield box and mask. None of them takes a multimodal dialogue as its interface. X-Chat inherits the parallel box and mask prediction of MaskDINO and admits box-only data as OpenSeeD and DaTaSeg do; the projection term of its consistency loss resembles that of DaTaSeg but compares the mask with the predicted box rather than an annotated one. Table 1 lists which of the fourteen tasks each model reports.

3 METHOD

X-Chat couples a multimodal LLM with a fine-grained perceptor so that a single instruction format covers seven task families, each in a segmentation form and a detection form. Figure 2 gives an overview.

3.1 FORMULATION

Inputs. An input consists of an image and a conversational instruction. Following X-SAM (Wang et al., 2026a), every entity the model has to localize is wrapped as a phrase `<p>···</p>` inside the instruction. Depending on the task, the phrase holds a category name (*generic* and *open-vocabulary* tasks), a referring expression (*referring* tasks), a question that requires reasoning (*reasoning* tasks), a phrase of the generated answer (*grounded conversation generation*, GCG), or a `<region>` placeholder that stands for a point, scribble, box, or mask prompt (*interactive* and *visual grounded*, VGD, tasks).

Outputs. The instruction also states the requested output form, for example "Please output the segmentation mask." or "Please output the bounding boxes." A detection task therefore differs from its segmentation counterpart only in the requested output, and each of the seven segmentation task families of X-SAM (Wang et al., 2026a) obtains a detection form: generic detection, open-vocabulary detection, referring expression comprehension (REC), reasoning detection, GCG with boxes, interactive detection, and VGD detection.

Unified formulation. The LLM answers in text and marks the perception request with a special `<PCT>` token. The phrase spans and the `<PCT>` token pass the request to the perceptor, which returns a mask and a box for every object query and scores every query against every phrase. The requested output form selects which of the two predictions is returned, so all fourteen tasks share the same model, decoding path, and weights.

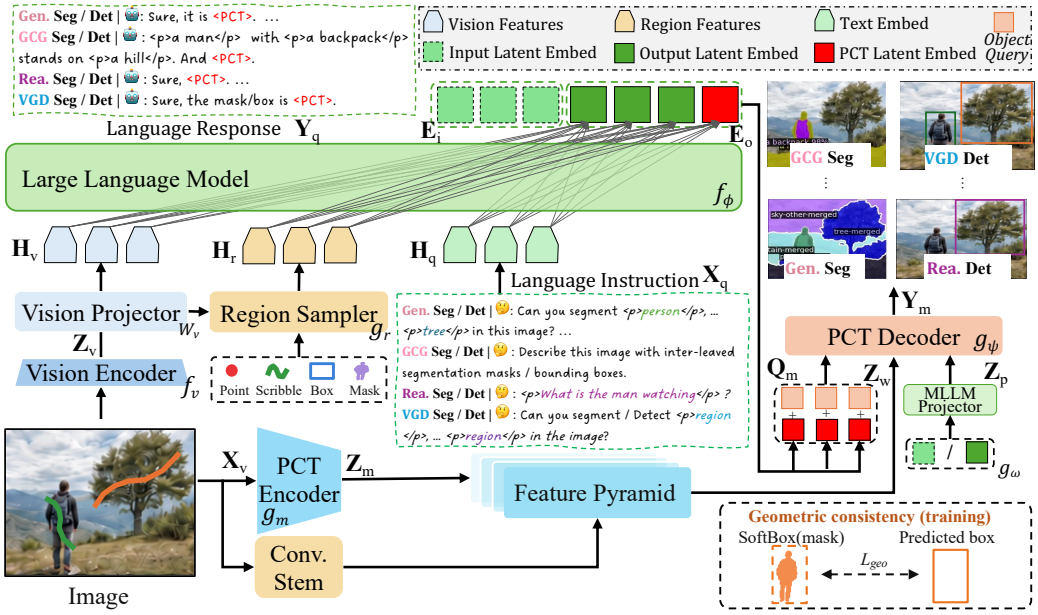

Figure 2: Overview of X-Chat. The multimodal LLM reads the image and a conversational instruction whose phrases carry category names, referring expressions, reasoning questions, or visual prompts (point, scribble, box, or mask) pooled by the region sampler. It answers in text and emits a <PCT> token, and the condition embeddings of the phrases and of <PCT> drive the perceptor (PCT encoder and PCT decoder in the figure), whose object queries each predict a mask and regress a box in parallel. The requested output form selects which of the two predictions is returned. During training, the geometric consistency loss couples each box with SoftBox of its mask.

3.2 Framework

Overview. X-Chat consists of a multimodal LLM, Qwen3-VL (Bai et al., 2025), a perceptor built from a DINOv3 ViT encoder (Siméoni et al., 2025) and a MaskDINO decoder (Li et al., 2023), and a region sampler for visual prompts. The LLM keeps its own vision encoder for image-level understanding, while the perceptor encodes the image separately for dense prediction.

Perceptor. The DINOv3 encoder produces a single-scale feature map, from which a simple feature pyramid in the style of ViTDet (Li et al., 2022b) builds multi-scale features. Because a plain ViT has a single coarse stride, a lightweight convolutional detail stem computes high-resolution maps directly from the pixels and adds them to the two finest pyramid levels through per-scale scalar gates initialized at zero. The MaskDINO decoder refines a set of object queries $\{q_i\}$; as in MaskDINO, each query predicts a mask $\hat{m}_i$ from the per-pixel embedding map and regresses a box $\hat{b}_i$ in parallel (Li et al., 2023).

Region sampler. For a visual prompt, the region sampler pools image features inside the prompted region and substitutes the pooled feature for the <region> placeholder, following X-SAM (Wang et al., 2026a). The prompt then enters the LLM as a phrase and reaches the perceptor through its condition embedding like any textual phrase.

Language-conditioned scoring. For each phrase k, the LLM hidden states over the phrase span are averaged into a condition embedding e_k, and the hidden state of the <PCT> token, e_{pct}, is added to every decoder query. The perceptor has no fixed label set: the score of query q_i for phrase k is the similarity between the decoded query and the normalized e_k, and a learned background embedding absorbs queries that match no phrase. The same scores drive query selection and rank both outputs of a query, so a category list, a referring expression, a reasoning question, and a visual prompt all select objects through one mechanism.

Box prediction with geometric consistency. Since $\hat{b}_i$ and $\hat{m}_i$ come from the same query state, a detection task needs no additional head. Parallel outputs alone, however, are free to dis-

agree, so X-Chat couples them with a geometric consistency loss on the matched queries of the final decoder layer. The loss first converts the mask into a differentiable box $\mathrm{SoftBox}(\hat{m}_i)$: the mask probabilities are max-projected onto each image axis, and the box edges are the expected positions of the first and last foreground location along each projection. It then combines the following terms,

$$\mathcal{L}_{\mathrm{geo}} = \lambda_1 \left\|\hat{b}_i - \mathrm{SoftBox}(\hat{m}_i)\right\|_1 + \lambda_2 \mathcal{L}_{\mathrm{GIoU}}\big(\hat{b}_i, \mathrm{SoftBox}(\hat{m}_i)\big) + \lambda_3 \mathcal{L}_{\mathrm{out}}(\hat{m}_i, \hat{b}_i) + \lambda_4 \mathcal{L}_{\mathrm{proj}}(\hat{m}_i, \hat{b}_i), \tag{1}$$

where $\mathcal{L}_{\mathrm{out}}$ penalizes mask probability outside a soft rasterization of $\hat{b}_i$, and $\mathcal{L}_{\mathrm{proj}}$ is a Dice loss between the axis projections of the mask and those of the rasterized box. $\mathcal{L}_{\mathrm{geo}}$ couples the supervision of the two outputs without exchanging features between them, and it adds no parameters. Section 4 compares this loss with explicit feature exchange between the box and mask predictions.

3.3 Training

Training has two stages; Table 6 in the appendix lists the hyperparameters.

Perceptor pre-training. The perceptor is first trained alone on COCO panoptic segmentation (Lin et al., 2014) with the classification, mask, and box losses of MaskDINO (Li et al., 2023) together with $\mathcal{L}_{\mathrm{geo}}$.

Joint multi-task fine-tuning. The pre-trained perceptor is then connected to the LLM and trained on general visual instruction data together with the training sets of ten of the fourteen tasks (Table 6). The open-vocabulary and interactive tasks, in both output forms, have no training set in this mixture. The vision encoder of the LLM stays frozen, the LLM is adapted with LoRA (Hu et al., 2021), and the perceptor is fine-tuned with a reduced learning rate on its encoder. Datasets are balanced by resampling, following X-SAM (Wang et al., 2026a). The objective is the autoregressive text loss plus the perceptor losses of the first stage. Samples that carry only box annotations contribute the classification and box losses and skip the mask and geometric consistency losses, so detection data enter training without being converted into pseudo masks.

4 Experiments

4.1 Tasks, Datasets, and Metrics

Segmentation. We evaluate the segmentation form of the seven task families on the benchmarks used by X-SAM (Wang et al., 2026a). *Generic* segmentation is measured on COCO (Lin et al., 2014) panoptic, semantic, and instance segmentation with panoptic quality (PQ), mean intersection over union (mIoU), and mask AP, and *open-vocabulary* segmentation on ADE20K (Zhou et al., 2017) (A150) with the same three metrics. *Referring* segmentation uses RefCOCO, RefCOCO+ (Yu et al., 2016), and RefCOCOg (Mao et al., 2016) with cumulative IoU (cIoU), and *reasoning* segmentation uses ReasonSeg (Lai et al., 2024) with per-sample mean IoU (gIoU). GCG follows the protocol of GLaMM (Rasheed et al., 2024) and reports mIoU and AP at IoU 0.5 (AP50) on its validation split. *Interactive* segmentation is measured on COCO-Interactive (Zhang et al., 2024c) with gIoU, and *VGD* segmentation on COCO-VGD (Wang et al., 2026a) with AP, each under point and box prompts.

Detection. The detection form is evaluated with box outputs. *Generic* detection reports box AP on COCO val2017, and *open-vocabulary* detection reports box AP on LVIS (Gupta et al., 2019) minival and the mean AP over the ODinW-13 datasets (Li et al., 2022a). REC reports the fraction of predictions with IoU at least 0.5 (Acc@0.5) on the validation splits of RefCOCO, RefCOCO+, and RefCOCOg. *Reasoning* detection uses RIO (Qu et al., 2023), whose common and uncommon splits are scored with AP50 as in the original benchmark. GCG with boxes is evaluated on Flickr30k Entities (Plummer et al., 2015) with mIoU and AP50. No benchmark exists for the *interactive* and *VGD* detection forms, so we build LVIS-INT and LVIS-VGD on LVIS images and report Acc@0.5 and AP, respectively, under point and box prompts; the appendix describes their construction and statistics (Table 7).

Table 2: Perceptor ablation. Encoder rows train the perceptor (encoder with a MaskDINO decoder) for 12 epochs under one fixed recipe with one seed and report the end-of-schedule COCO val2017 panoptic and instance metrics. Adaptation rows keep the DINOv3-S+ encoder and vary how its single-scale features are turned into a feature pyramid, each arm trained once under the same recipe as the encoder rows.

Perceptor	PQ	mIoU	Mask AP	Box AP
Encoder				
Swin-T	50.6	59.6	42.0	44.5
DINOv3-S+	55.0	65.3	44.7	47.8
DINOv3-L	59.8	69.7	50.7	55.5
Adaptation structure				
ViTDet-style simple pyramid	55.0	65.8	44.9	48.8
Simple pyramid + detail stem	56.1	65.9	46.7	50.4

Table 3: Box-only detection data in joint training. All arms train on COCO panoptic for the same number of steps with one seed; arms 2 and 3 spend most steps on box-only batches from the full Objects365 training set. Arm 3 uses the same images and boxes as arm 2 with every category name replaced by "object". ADE20K (A150) measures open-vocabulary segmentation; PQ^{th}_{O365} is the thing PQ on the ADE20K categories that Objects365 names and COCO does not, under a lenient (11 categories) and a strict (20 categories) matching of category names.

#	Objects365		ADE20K (A150)		COCO
	Boxes	Names	PQ/mIoU/AP	PQ^{th}_{O365} len./str.	PQ
1			18.3/21.2/11.7	4.5/7.3	56.4
2	✓	✓	21.1/25.2/18.7	31.4/29.6	55.0
3	✓		15.3/17.2/10.9	0.0/1.8	54.5

Table 4: Geometric consistency for box–mask coupling. Both settings train the perceptor (DINOv3-S+ encoder, MaskDINO decoder) from the same initialization on COCO panoptic segmentation for 12 epochs with one seed, and are evaluated on COCO val2017. Row 1 uses parallel box and mask heads, and row 2 trains them with the geometric consistency loss (Geo.), which adds no parameters. Feature-exchange and gating variants are in Table 8.

#	Geo.	Det.	Seg.
		Box AP	Mask AP/PQ/mIoU
1		47.5	45.7/55.8/66.1
2	✓	49.0	45.5/55.7/66.0

4.2 Implementation Details

The main model, X-Chat, combines Qwen3-VL-4B (Bai et al., 2025) with a DINOv3-L perceptor (Siméoni et al., 2025) and is trained with the two stages of Section 3.3; Table 6 lists its hyperparameters. All benchmark results in Table 5 come from one set of weights, without task-specific fine-tuning. The perceptor ablations (Tables 2 and 4) train the perceptor alone on COCO panoptic segmentation with a fixed schedule, and the joint-training ablation (Table 3) fine-tunes a separate X-Chat configuration, Qwen3-VL-2B with a DINOv3-S+ perceptor, on COCO panoptic with or without Objects365 boxes. Every ablation arm is trained once, so small differences between arms are not separated from run-to-run variation.

4.3 Ablation Studies

Fine-grained branch. Table 2 compares perceptor encoders under the same pre-training schedule on COCO. DINOv3-L reaches 59.8 PQ and 55.5 box AP, ahead of DINOv3-S+ (55.0 PQ, 47.8 box AP) and Swin-T (50.6 PQ, 44.5 box AP). The adaptation structures that turn the single-scale ViT output into a feature pyramid are compared in the lower block of the same table. We adopt DINOv3-L with a simple feature pyramid and the detail stem as the perceptor of X-Chat.

Box–mask coupling. Table 4 compares parallel box and mask heads (row 1) with the same heads trained with the geometric consistency loss (row 2). The geometric consistency loss raises box AP from 47.5 to 49.0 (+1.5) with the same parameter count as row 1, while mask AP and PQ stay within 0.2 of it. Explicit feature exchange on top of this setting does not add to the gain (Table 8): relative

to the geometric consistency loss alone (row 2), mask-to-box injection lowers box AP by 1.3, box-to-mask injection changes mask AP by -0.05, and the gated bidirectional variant differs by +0.02 box AP and +0.17 mask AP. These results support geometric consistency over parallel heads and show no benefit from explicit exchange, but only for this schedule: all arms were still improving at its end, with box AP rising by more than the spread between arms over the last evaluation interval. We adopt the geometric consistency loss, which adds no parameters and no inference-time module.

Detection data in joint training. Table 3 adds box-only Objects365 (Shao et al., 2019) data to joint fine-tuning on COCO panoptic and evaluates open-vocabulary segmentation on ADE20K. Adding the Objects365 boxes with their category names (row 2) raises ADE PQ from 18.3 to 21.1, whereas the same images and boxes with every name replaced by "object" (row 3) lower it to 15.3. These runs use the full Objects365 training set; an earlier round of the same three arms on a 12,000-image Objects365 subset for 4,000 steps is a separate result set and is not compared here. The gain concentrates on ADE categories that only Objects365 names: their thing PQ rises from 4.5 to 31.4 with names and falls to 0.0 without them under the lenient category matching of Table 3, and moves from 7.3 to 29.6 and 1.8 under the strict one. The lenient O365-only group has only 11 categories; the wider group of ADE categories that COCO does not name shows the same ordering, with thing PQ of 3.7, 17.4, and 0.0 for rows 1, 2, and 3 under the lenient matching (25 categories) and 5.0, 13.2, and 2.3 under the strict one (60 categories). Box supervision therefore helps open-vocabulary segmentation mainly through the category names it brings into the condition embeddings, not through the boxes alone. The comparison has limits: the instructions of rows 2 and 3 differ in form as well as in vocabulary, the box-only batches reduce the number of segmentation steps, and in-domain COCO PQ drops from 56.4 to 55.0.

4.4 Benchmark Results

Detection. The upper part of Table 5 reports the detection form: X-Chat produces boxes for all seven task families with a single model. On COCO it reaches 55.8 box AP, above the 39.0 AP that PaDT Pro-7B reports on COCO val2017 (Su et al., 2025). On REC it obtains 94.4, 90.2, and 94.1 Acc@0.5 on the RefCOCO, RefCOCO+, and RefCOCOg validation splits, compared with 89.5 and 83.9 for Groma (Ma et al., 2024) on RefCOCO and RefCOCO+. On the reasoning benchmark RIO it reaches 79.4 and 47.1 AP50 on the common and uncommon splits, against 49.1 and 22.0 for TOIST as reported by Qu et al. (2023). For open-vocabulary detection, X-Chat obtains 44.1 AP on LVIS minival and 52.5 mean AP on ODinW-13; the LVIS result is not zero-shot, because LVIS training images are part of the joint fine-tuning data. Grounded captioning with boxes has been evaluated before: UniTAB (Yang et al., 2022) reports two F1 scores, the harmonic mean of precision and recall, over all predicted object words ($F1_{all}$) and over correctly predicted object words only ($F1_{loc}$), on the Flickr30k Entities test split, LLaVA-Grounding (Zhang et al., 2024a) reports an F1 score on its own Grounding-Bench, and REF-VLM (Tai et al., 2025) reports zero-shot mIoU and AP50 on the Flickr30k test split. None of them reports GCG with boxes on the validation split used here, and no published method reports interactive detection or VGD detection. On these forms X-Chat reaches 71.7 mIoU on the GCG-Flickr validation split, 38.2 and 88.5 Acc@0.5 on LVIS-INT under point and box prompts, and 46.1 and 54.0 AP on LVIS-VGD under point and box prompts. For interactive detection we therefore ran Qwen3-VL-4B zero-shot on LVIS-INT with a fixed prompt template: with point prompts it reaches 4.3 Acc@0.5 against 38.2 for X-Chat, and with box prompts 97.3 against 88.5. Returning the box of the prompt itself already scores 4.2 and 98.5, above every model on box prompts, so only the point-prompt column separates the models. We report the 4B variant because it is the LLM inside X-Chat; the comparison still mixes the perceptor with training exposure, since X-Chat was trained on the same four prompt types on LVIS train and Qwen3-VL was not, and 11.6% of the 4B point-prompt outputs fail to parse and count as zero IoU. For VGD detection we ran Rex-Omni-3B with box prompts on LVIS-VGD: it outputs no confidence scores, so its 41.4 AP ranks boxes in output order and is a reference value rather than a comparison; its own protocol scores F1 at IoU 0.5, which we have not computed for X-Chat.

Segmentation. The lower part of Table 5 compares the segmentation form with X-SAM, which shares its task formulation but outputs no boxes. X-Chat reaches 57.4 PQ on COCO and 27.4 PQ on A150, compared with 54.7 and 20.9 for X-SAM (Wang et al., 2026a). On referring and reasoning segmentation the two models are close: X-Chat obtains 84.6, 79.2, and 85.45 cIoU on RefCOCO, RefCOCO+, and RefCOCOg against 85.1, 78.0, and 83.8, and 58.6 and 57.6 gIoU on ReasonSeg validation and test against 56.6 and 57.8 for X-SAM (Wang et al., 2026a). X-Chat is also higher

Table 5: Comparison with specialist models and MLLM generalists on the detection form (top) and the segmentation form (bottom) of the seven task families. "–" denotes a result not available in the sources we hold. Among the MLLM generalists, the best result for each metric is in **bold** and the second best, marked only where at least three results are compared, is underlined. Specialists are not ranked, nor are results whose metric, split, or protocol is not established to match the column: the † cell, REF-VLM on referring detection, X-SAM on the ReasonSeg test split, Qwen3-VL-4B with box prompts, and Rex-Omni-3B. On COCO-Interactive, the X-Chat value is gIoU, the mean of per-sample IoU, which we read as the mIoU of X-SAM, whose released evaluator computes the same quantity. † GLaMM does not name the metric of its referring segmentation table, and this row is its model fine-tuned on RefCOCO, RefCOCO+, and RefCOCOg. X-Chat uses a single set of weights for all tasks. Qwen3-VL-4B is run zero-shot on LVIS-INT with one frozen prompt template. Rex-Omni-3B is evaluated on LVIS-VGD with box prompts on a fixed set of prompt instances without category names; it outputs no confidence scores, so its AP ranks boxes in output order and is a reference value only.

Method	*Detection*						
	Gen. COCO AP	***OV*** LVIS / ODinW AP	***Ref.*** RefCOCO/+/g val Acc@0.5	***Rea.*** RIO com./uncom. AP50	***GCG*** Flickr mIoU/AP50	***Int.*** LVIS-INT pt./box Acc@0.5	***VGD*** LVIS-VGD pt./box AP
Specialists							
MaskDINO-R50 (Li et al., 2023)	51.7	–	–	–	–	–	–
Grounding DINO (zero-shot) (Liu et al., 2024b)	52.5	33.9 / 26.1	–	–	–	–	–
MDETR	–	–	–	48.6 / 24.2	–	–	–
TOIST	–	–	–	49.1 / 22.0	–	–	–
MLLM Generalists							
VisionLLM v2 (Wu et al., 2024)	**56.7**	–	87.9 / 77.6 / –	–	–	–	–
Groma-7B (Ma et al., 2024)	–	–	89.5 / 83.9 / –	–	–	–	–
REF-VLM (Tai et al., 2025)	–	–	90.8 / – / 87.0	–	–	–	–
PaDT Pro-7B (Su et al., 2025)	39.0	–	–	–	–	–	–
Qwen3-VL-4B (Bai et al., 2025)	–	–	–	–	–	4.3 / 97.3	–
Rex-Omni-3B	–	–	–	–	–	–	– / 41.4
X-Chat	55.8	44.1 / 52.5	**94.4** / **90.2** / 94.1	79.4 / 47.1	71.7 / 19.9	**38.2** / 88.5	46.1 / 54.0

Method	*Segmentation*						
	Gen. COCO PQ/mIoU/AP	***OV*** A150 PQ/mIoU/AP	***Ref.*** RefCOCO/+/g val cIoU	***Rea.*** ReasonSeg val/test gIoU	***GCG*** val mIoU/AP50	***Int.*** COCO-Int. pt./box mIoU	***VGD*** COCO-VGD pt./box AP
Specialists							
Mask2Former-L (Cheng et al., 2022)	57.8 / – / –	–	–	–	–	–	–
MaskDINO-L (Li et al., 2023)	59.4 / – / 54.5	–	–	–	–	–	–
ODISE	–	22.6 / – / 14.4	–	–	–	–	–
SAM-L (Kirillov et al., 2023)	–	–	–	–	–	– / 76.6	–
MLLM Generalists							
LISA-7B (Lai et al., 2024)	–	–	74.1 / 62.4 / 66.4	44.4 / –	–	–	–
GLaMM (Rasheed et al., 2024)	–	–	79.5 / 72.6 / 74.2†	–	66.3 / 30.8	–	–
PixelLM (Ren et al., 2024)	–	–	73.0 / 66.3 / 69.3	–	–	–	–
PSALM (Zhang et al., 2024c)	55.9 / 66.6 / 45.7	13.7 / 18.2 / 9.0	83.6 / 72.9 / 73.8	–	–	64.3 / 67.3	–
VisionLLM v2 (Wu et al., 2024)	– / – / 47.8	–	76.6 / – / –	–	–	– / 59.1	–
MVP-LM (Xiang et al., 2025)	56.1 / 66.8 / –	19.4 / 20.5 / –	83.6 / 73.9 / 75.1	–	–	–	–
UFO-8B (Tang et al., 2025)	–	–	–	– / 55.4	–	–	–
X-SAM (Wang et al., 2026a)	54.7 / 66.5 / –	20.9 / – / –	**85.1** / 78.0 / 83.8	56.6 / 57.8	69.4 / 33.2	**65.4** / **69.6**	47.9 / 49.5
X2SAM (Wang et al., 2026b)	54.1 / 64.8 / 45.8	**31.2** / **38.2** / 20.2	84.0 / 78.4 / 81.9	**71.1** / **68.5**	67.1 / –	–	45.9 / 48.5
X-Chat	**57.4** / **69.0** / **51.7**	27.4 / 33.2 / **21.3**	84.6 / **79.2** / **85.45**	58.6 / 57.6	**70.05** / **35.7**	46.6 / 54.2	**49.4** / **52.8**

on GCG (70.05 mIoU and 35.7 AP50 against 69.4 and 33.2) and COCO-VGD (49.4 and 52.8 AP under point and box prompts against 47.9 and 49.5). Interactive segmentation is the exception: X-Chat reaches 46.6 and 54.2 gIoU under point and box prompts, below the 65.4 and 69.6 mIoU of X-SAM (Wang et al., 2026a). X-SAM's released evaluator reports interactive segmentation only as cIoU and gIoU, the metric computed for X-Chat, so we read its mIoU as gIoU and the gap as one on the same metric. X-Chat thus exceeds the segmentation-only X-SAM on generic, open-vocabulary, GCG, and VGD segmentation, is close to it on referring and reasoning segmentation, and is below it on interactive segmentation, while also producing boxes; since the two models differ in LLM and perceptor, this comparison does not isolate the effect of adding the detection form.

5 Discussion

Conclusion. X-Chat takes multimodal dialogue as its perception interface and covers the segmentation and detection forms of seven task families, fourteen tasks in all, with one model. Each object query of its perceptor predicts a mask and regresses a box in parallel, and a geometric consistency loss couples the two, so the detection form of a task needs no head of its own and box-only data such as Objects365 enter joint training directly.

AI USE STATEMENT

We used generative artificial intelligence (AI) tools to assist in implementing the experiment code and the data processing. We used a large language model (Claude, Anthropic) to draft and edit the manuscript text and to typeset its tables from our result files. The research ideas are the authors' own. Every number in the paper was checked against the evaluation logs and configuration files it cites, and every statement about prior work against our reading notes of that work. All authors reviewed the AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All training and evaluation data come from existing datasets, cited where they are used (sections A.2 and 4.1). LVIS-VGD and LVIS-INT contain no new images and no new human annotation: their images are COCO images, their instances are LVIS annotations, and their point, scribble, box, and mask prompts are generated automatically from the instance masks (section A.3). The benchmarks therefore inherit the COCO images and the LVIS category vocabulary.

REPRODUCIBILITY STATEMENT

Section 3.3 describes the training procedure and section 4.2 the implementation. Section A.1 lists the training hyperparameters in table 6, and section A.2 names every dataset in the joint fine-tuning mixture. Section A.3 gives the construction rules and statistics of LVIS-VGD and LVIS-INT (table 7).

REFERENCES

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

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 1280–1289. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00135.

Xiuye Gu, Yin Cui, Jonathan Huang, Abdullah Rashwan, Xuan Yang, Xingyi Zhou, Golnaz Ghiasi, Weicheng Kuo, Huizhong Chen, Liang-Chieh Chen, and David A. Ross. Dataseg: Taming a universal multi-dataset multi-task segmentation model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d4eed238cf5807c6b75face996302892-Abstract-Conference.html.

Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 5356–5364. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00550. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Gupta_LVIS_A_Dataset_for_Large_Vocabulary_Instance_Segmentation_CVPR_2019_paper.html.

Xiaoyang Han, Jianhua Li, Kewang Deng, Zukai Chen, Xuanke Shi, Sihan Wang, Boxuan Li, Linyan Wang, Siyi Xie, Xin You, Jinsheng Quan, Zhongang Cai, Haiwen Diao, Ziwei Liu, Lei Yang, Dahua Lin, and Quan Wang. Vision as unified multimodal generation, 2026.

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, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3992–4003. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00371.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: reasoning segmentation via large language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9579–9589. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00915. URL https://doi.org/10.1109/CVPR52733.2024.00915.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask DINO: towards A unified transformer-based framework for object detection and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 3041–3050. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00297.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965. IEEE, 2022a. doi: 10.1109/cvpr52688.2022.01069.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Computer Vision – ECCV 2022*, pp. 280–296. Springer Nature Switzerland, 2022b. doi: 10.1007/978-3-031-20077-9_17.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a. doi: 10.1109/cvpr52733.2024.02484.

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 Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII*, volume 15105 of *Lecture Notes in Computer Science*, pp. 38–55. Springer, 2024b. doi: 10.1007/978-3-031-72970-6_3.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pp. 417–435. Springer, 2024. doi: 10.1007/978-3-031-72658-3_24. URL https://doi.org/10.1007/978-3-031-72658-3_24.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 11–20. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.9.

Zhixiong Nan, Xianghong Li, Tao Xiang, and Jifeng Dai. Di-maskdino: A joint object detection and instance segmentation model. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6f1346bac8b02f76a631400e2799b24b-Abstract-Conference.html.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2641–2649. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.303.

Mengxue Qu, Yu Wu, Wu Liu, Xiaodan Liang, Jingkuan Song, Yao Zhao, and Yunchao Wei. RIO: A benchmark for reasoning intention-oriented objects in open environments. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/8644353f7d307baaf29bc1e56fe8e0ec-Abstract-Datasets_and_Benchmarks.html.

Hanoona Abdul Rasheed, Muhammad Maaz, Sahal Shaji Mullappilly, Abdelrahman M. Shaker, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Eric P. Xing, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Glamm: Pixel grounding large multimodal model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13009–13018. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01236. URL https://doi.org/10.1109/CVPR52733.2024.01236.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26364–26373. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02491. URL https://doi.org/10.1109/CVPR52733.2024.02491.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 8429–8438. IEEE, 2019. doi: 10.1109/ICCV.2019.00852.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

Yongyi Su, Haojie Zhang, Shijie Li, Nanqing Liu, Jingyi Liao, Junyi Pan, Yuan Liu, Xiaofen Xing, Chong Sun, Chen Li, Nancy F. Chen, Shuicheng Yan, Xulei Yang, and Xun Xu. Patch-as-decodable-token: Towards unified multi-modal vision tasks in mllms, 2025.

Yan Tai, Luhao Zhu, Yunan Ding, Yiying Dong, Guangtao Zhai, Xiaohong Liu, and Guodong Guo. Ref-vlm: Triplet-based referring paradigm for unified visual decoding, 2025.

Hao Tang, Chen-Wei Xie, Haiyang Wang, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. UFO: A unified approach to fine-grained visual perception via open-ended language interface. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/78a241e1d2ff2366af8cc9a459875a01-Abstract-Conference.html.

Guanqun Wang, Xinyu Wei, Jiaming Liu, Ray Zhang, Yichi Zhang, Kevin Zhang, Maurice Chong, and Shanghang Zhang. Mr-mllm: Mutual reinforcement of multimodal comprehension and vision perception, 2024.

Hao Wang, Limeng Qiao, Zequn Jie, Zhijian Huang, Chengjian Feng, Qingfang Zheng, Lin Ma, Xiangyuan Lan, and Xiaodan Liang. X-SAM: from segment anything to any segmentation. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 26187–26196. AAAI Press, 2026a. doi: 10.1609/AAAI.V40I31.39822.

Hao Wang, Limeng Qiao, Chi Zhang, Lin Ma, Guanglu Wan, Xiangyuan Lan, and Xiaodan Liang. X2sam: Any segmentation in images and videos, 2026b.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, Jan Kautz, Lei Zhang, and Zhiding Yu. Locateanything: Fast and high-quality vision-language grounding with parallel box decoding, 2026c.

Xiaolong Wang, Lixiang Ru, Ziyuan Huang, Kaixiang Ji, DanDan Zheng, Jingdong Chen, and Jun Zhou. ARGenSeg: Image segmentation with autoregressive image generation model. In *Advances in Neural Information Processing Systems 38*, NeurIPS 2025, pp. 32117–32141. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025. doi: 10.52202/085713-0954.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Ping Luo, Yu Qiao, and Jifeng Dai. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/81a60d18e010b27b36cd465c6604b915-Abstract-Conference.html.

Wentao Xiang, Haoxian Tan, Yujie Zhong, Cong Wei, Dengjie Li, and Yujiu Yang. Advancing visual large language model for multi-granular versatile perception. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 22153–22164. IEEE, 2025. doi: 10.1109/ICCV51701.2025.02057. URL https://doi.org/10.1109/ICCV51701.2025.02057.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. UniTAB: Unifying text and box outputs for grounded vision-language modeling. In *Computer Vision – ECCV 2022*, pp. 521–539. Springer Nature Switzerland, 2022. ISBN 9783031200595. doi: 10.1007/978-3-031-20059-5_30.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pp. 69–85. Springer, 2016. doi: 10.1007/978-3-319-46475-6_5.

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, 2022a.

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 *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 1020–1031. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00100.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, and Jainwei Yang. LLaVA-Grounding: Grounded visual chat with large multimodal models. In *Computer Vision – ECCV 2024*, pp. 19–35. Springer Nature Switzerland, 2024a. ISBN 9783031727757. doi: 10.1007/978-3-031-72775-7_2.

Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022b. URL http://papers.nips.cc/paper_files/paper/2022/hash/ea370419760b421ce12e3082eb2ae1a8-Abstract-Conference.html.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation, 2024b.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. PSALM: pixelwise segmentation with large multi-modal model. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXIV*, volume 15092 of *Lecture Notes in Computer Science*, pp. 74–91. Springer, 2024c. doi: 10.1007/978-3-031-72754-2_5.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5122–5130. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.544.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 15116–15127. IEEE, 2023a. doi: 10.1109/CVPR52729.2023.01451. URL https://doi.org/10.1109/CVPR52729.2023.01451.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/3ef61f7e4afacf9a2c5b71c726172b86-Abstract-Conference.html.

Table 6: Training hyperparameters of the two stages (Section 3.3). In joint fine-tuning, the vision encoder of the multimodal model is frozen and LoRA (Hu et al., 2021) is applied to its LLM.

	Perceptor pre-training	**Joint multi-task fine-tuning**
Training data	COCO panoptic train2017	LLaVA-1.5 mix665k + 17 task training sets
Epochs	12	1
Batch size per device	2	4
Optimizer	AdamW, $\beta = (0.9, 0.999)$	AdamW, $\beta = (0.9, 0.999)$
Learning rate	4×10^{-4}	2×10^{-4}
LR multiplier, perceptor encoder	0.1	0.1
Weight decay	0.05	0.05
Warmup	linear, 500 iterations	linear, first 3% of training
LR decay	$\times 0.1$ at epochs 11 and 12	cosine to 0
Gradient clipping (max norm)	0.01	1
Precision	bf16	bf16
Perceptor input	1024×1024 crop, scale 0.1–2.0, flip	1024×1024
LoRA rank / α / dropout	—	128 / 256 / 0.05
Seed	1024	1024

A Additional Details

A.1 Training Details

Table 6 lists the hyperparameters of the two training stages of Section 3.3.

In perceptor pre-training, the MaskDINO decoder (Li et al., 2023) has 9 layers and 300 object queries. Its loss weights are 4 for classification, 5 for each of the mask and Dice losses, and 5 and 2 for the box and GIoU losses.

A.2 Training Data

The joint fine-tuning mixture combines the LLaVA-1.5 mix665k instruction data (Liu et al., 2024a) with 17 task training sets. For the segmentation forms, the task sets are COCO panoptic (Lin et al., 2014) for generic segmentation; RefCOCO, RefCOCO+ (Yu et al., 2016), and RefCOCOg (Mao et al., 2016) for referring segmentation; the ReasonSeg training data of LISA (Lai et al., 2024) for reasoning segmentation; four GCG sets, GranD-f (Rasheed et al., 2024), RefCOCOg-GCG, OpenPSG-GCG, and Flickr-GCG, for GCG segmentation; and COCO-VGD (Wang et al., 2026a) for VGD segmentation. For the detection forms, they are Objects365 v1 (Shao et al., 2019) for generic detection; the same three RefCOCO datasets for referring detection; RIO (Qu et al., 2023) for reasoning detection; the grounded visual chat data of LLaVA-Grounding (Zhang et al., 2024a) for GCG detection; and the training split of LVIS-VGD, built on LVIS (Gupta et al., 2019), for VGD detection.

A.3 LVIS-VGD and LVIS-INT

Table 7 gives the statistics of the two benchmarks, which share their images and instances.

The validation images are the LVIS MiniVal images from COCO val2017, not a larger LVIS validation set in which 15000 of 19809 images come from COCO train2017 and would overlap with the training images. Instances of categories that an image marks as not exhaustively annotated are removed, rare-category instances are removed from the training split only, and images left without instances are dropped. A point, a scribble, a box, and a mask prompt are generated for every connected component of an instance mask with an area larger than 5.

Table 7: Statistics of LVIS-VGD and LVIS-INT. The two benchmarks share their images and instances: the validation images are the LVIS MiniVal images from COCO val2017, and the training images come from COCO train2017. Both use the 1203-category LVIS vocabulary; the category column counts categories with at least one instance. The prompt columns count instances whose prompt of that type is non-empty; an evaluation sample is one image for LVIS-VGD and one instance for LVIS-INT. The validation split contains 65 rare categories with 169 instances that never appear in training.

				Prompted instances / evaluation samples			
	Images	**Instances**	**Categories**	**Point**	**Scribble**	**Box**	**Mask**
Validation (shared)	4662	36213	787	35991	35572	35967	36211
Training (shared)	97292	831856	866	824302	811334	823568	831688
LVIS-VGD evaluation samples	–	–	–	4661	4660	4661	4662
LVIS-INT evaluation samples	–	–	–	35991	35572	35967	36211

Table 8: Full ablation on box–mask coupling. All arms train the perceptor (DINOv3-S+ encoder, MaskDINO decoder) from the same initialization on COCO panoptic segmentation for 12 epochs with one seed, and are evaluated on COCO val2017. Geo.: geometric consistency loss; M→B / B→M: mask-to-box / box-to-mask feature injection into the final prediction head; Gate: confidence gate with interaction warm-up. Rows 1 and 2 are the two settings of Table 4; the adopted setting is highlighted.

#	**Geo.**	**M→B**	**B→M**	**Gate**	**Extra** Params	**Det.** Box AP	**Seg.** Mask AP/PQ/mIoU
1					0	47.5	45.7/55.8/66.1
2	✓				0	49.0	45.5/55.7/66.0
3	✓	✓			65.8K	47.7	45.4/55.2/66.0
4	✓		✓		65.8K	48.8	45.5/55.5/65.7
5	✓	✓	✓		131.6K	48.0	45.5/55.4/66.0
6	✓	✓	✓	✓	131.7K	49.1	45.7/55.5/65.7

A.4 Box–Mask Coupling Ablation

Table 8 extends Table 4 with explicit exchange between the box and mask predictions on top of the geometric consistency loss: mask-to-box and box-to-mask feature injection into the final prediction head (rows 3 and 4), both directions together (row 5), and a confidence gate on the injection (row 6). None of these variants improves on row 2 by more than 0.2 box AP or mask AP, while each adds between 65.8K and 131.7K parameters. All arms were still improving at the end of the 12-epoch schedule, so this comparison holds for that schedule only.

A.5 Mask Supervision for Box-Only Data

Table 9 reports an earlier round of the Objects365 joint-training ablation, trained for 4,000 steps on a 12,000-image subset of Objects365 with one seed. It asks whether box-only samples gain from mask supervision derived from their boxes. This round is a separate result set from the full-data round of Table 3, and no number is compared across the two. The reference arm (row 2) adds the Objects365 boxes without any mask loss on box-only samples; it reaches 22.2 ADE20K PQ, whereas pseudo-mask distillation from an exponential-moving-average teacher (rows 4–6) reaches 21.4 to 21.8 across three settings of its strength and supervised outputs, and three box-projection weak losses (rows 7–9) reach 21.4 to 21.9. In-domain COCO PQ stays within 0.4 of the reference arm for all six variants. A same-seed rerun of the reference arm differs from it by 0.8 ADE20K PQ, so these single runs give no evidence that either form of derived mask supervision helps open-vocabulary segmentation.

Table 9: Mask supervision for box-only samples, earlier Objects365 round. All arms fine-tune on COCO panoptic segmentation for 4,000 steps with one seed; rows 2–9 add the boxes of a 12,000-image Objects365 subset, and row 1 does not. Row 2 applies no mask loss to box-only samples; rows 4–6 add pseudo-mask distillation from an exponential-moving-average teacher, and rows 7–9 add box-projection weak losses on the final output. Row 3 reruns row 2 with the same seed; only its ADE20K PQ has been analyzed. Evaluation on ADE20K (A150) panoptic val and COCO panoptic val2017. These runs are a separate result set from Table 3 and are not compared with it.

#	**Mask supervision of box-only samples**	**ADE20K (A150)**			**COCO**		
		PQ	mIoU	AP	PQ	mIoU	AP
1	No Objects365 data	18.9	22.0	12.9	56.4	67.7	47.4
2	None	22.2	26.9	16.1	56.4	68.1	47.6
3	None (rerun, same seed)	23.0	–	–	–	–	–
4	Distillation, weight 1, final and auxiliary outputs	21.7	26.7	16.0	56.6	68.4	47.6
5	Distillation, weight 0.2, final and auxiliary outputs	21.8	26.4	16.2	56.6	68.2	47.6
6	Distillation, weight 1, final output only	21.4	26.7	16.1	56.4	68.1	47.6
7	Box projection	21.9	26.1	16.4	56.6	67.8	47.6
8	Box projection, local consistency	21.5	26.3	16.1	56.7	68.1	47.6
9	Box projection, local consistency, fractional boundary target	21.4	25.7	16.2	56.8	68.1	47.8

A.6 More Visualization Results

Figures 3 to 9 show additional qualitative results of X-Chat, one figure per task family, with detection examples on the left and segmentation examples on the right. In each example, the answer marks the returned boxes or masks with a <PCT> token.

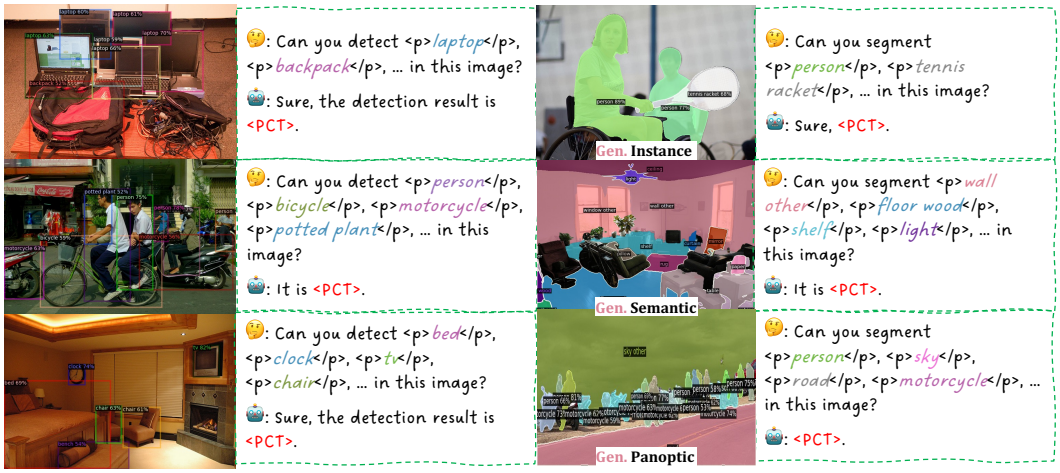

Figure 3: Visualization results of generic detection and generic segmentation; the segmentation examples cover instance, semantic, and panoptic segmentation.

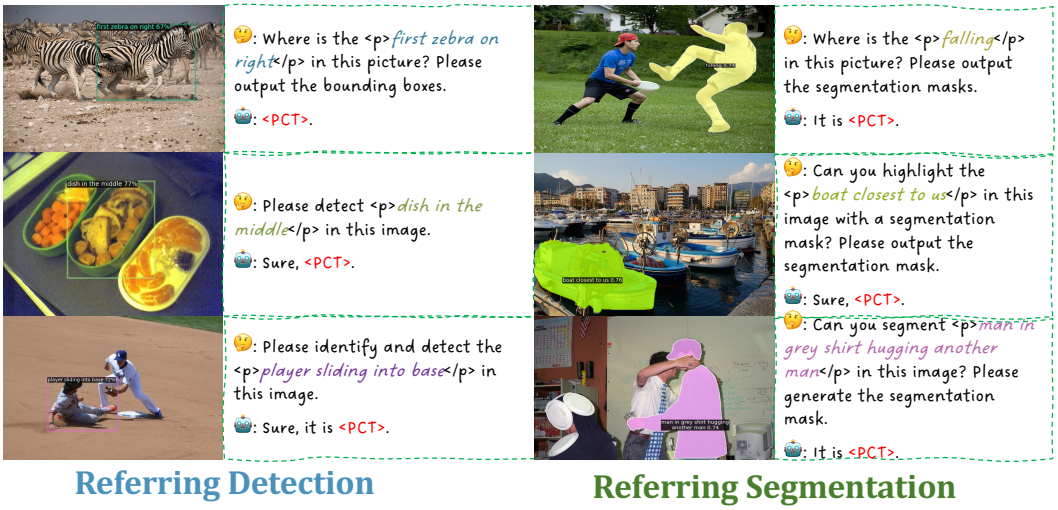

Figure 4: Visualization results of referring detection and referring segmentation.

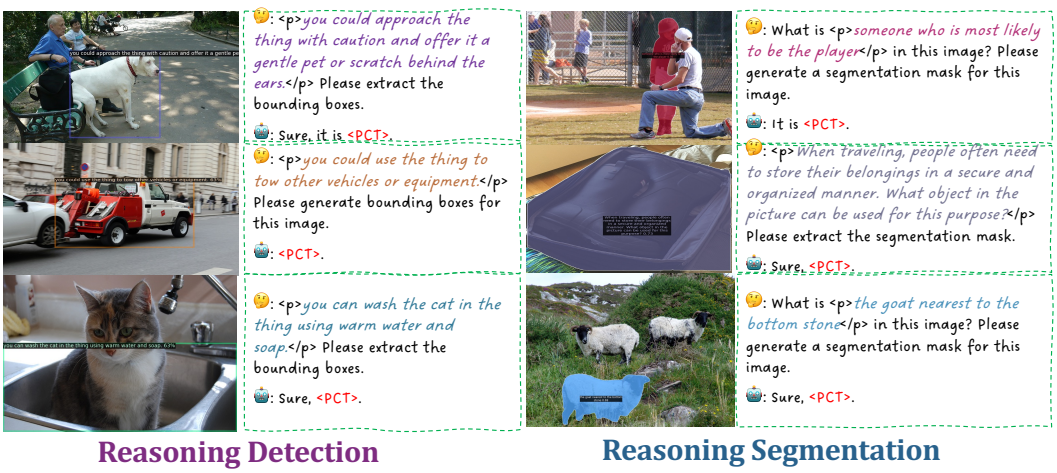

Figure 5: Visualization results of reasoning detection and reasoning segmentation.

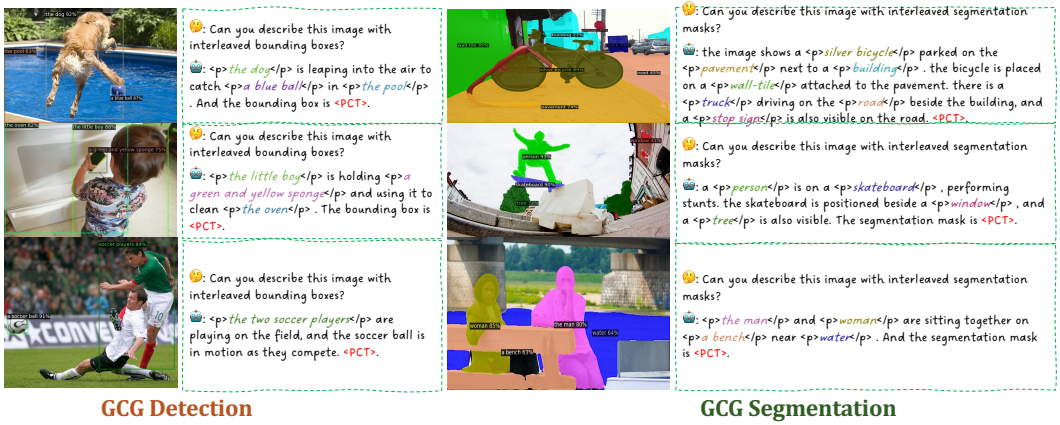

Figure 6: Visualization results of grounded conversation generation (GCG) with boxes and with masks.

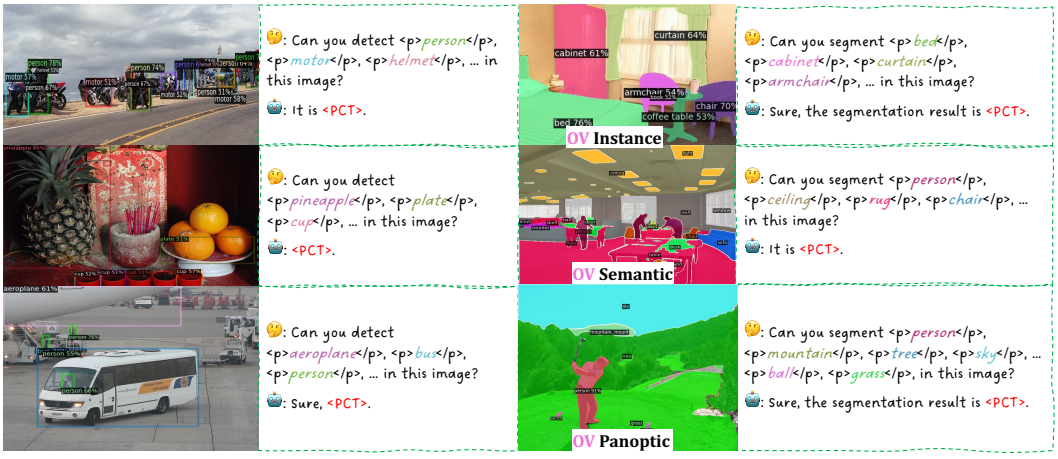

Figure 7: Visualization results of open-vocabulary (OV) detection and segmentation; the segmentation examples cover instance, semantic, and panoptic segmentation.

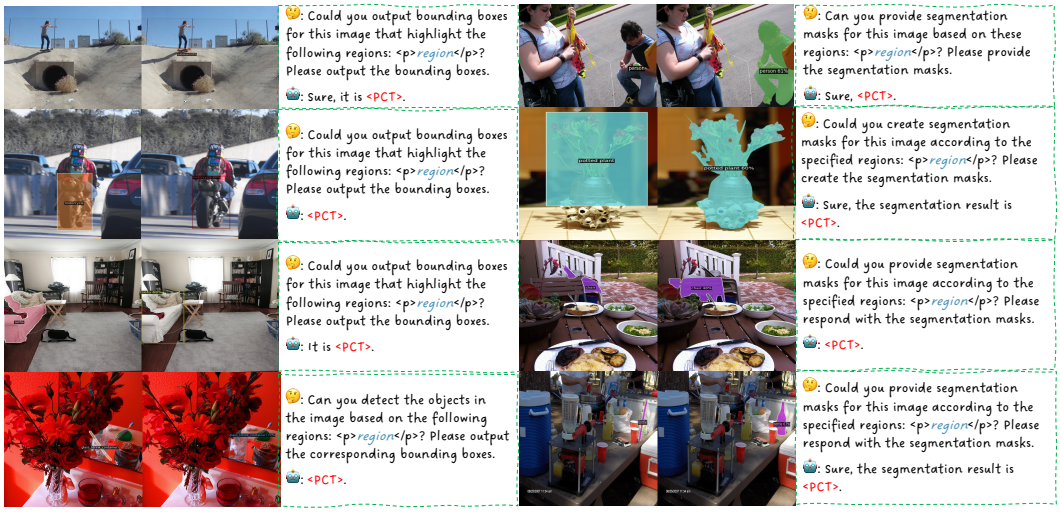

Figure 8: Visualization results of interactive detection and interactive segmentation, where a visual prompt marks the target region.



Figure 9: Visualization results of visual grounded (VGD) detection and segmentation, where visual prompts select every instance of the prompted categories.