SKETCH-OMNI: INSTANCE-LEVEL SKETCH-BASED VISUAL PROMPTING FOR PERCEPTION MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Promptable perception, from GroundingDINO to Rex-Omni, puts one model behind one prompt interface and returns coordinates for any localisation task. Every model in that line was built for photographs; none has been given a sketch. We present **Sketch-Omni**, to our knowledge the first perception MLLM built to work *on* sketches and *from* them. On a sketch, the sketch is the image, and text prompts detect, point, resolve object and spatial referring expressions, and ground affordances in it. From a sketch, the sketch is the prompt, and a drawing of one object localises that object in a photograph, even among several of the same category, a query that neither a category name nor a photo exemplar can pose. Two obstacles stand in the way. A photo-trained model's attention is tuned to colour and texture, which a sketch does not have, and no dataset pairs a drawing with the one instance it depicts among others of its kind. We build the training data instead of collecting it. An edge detector turns every labelled photograph into a sketch-like drawing that keeps the photograph's labels, and two constructions produce sketch-to-photo instance pairs from existing data. On top of this, *photo-guided relation alignment* passes each photograph and its edge sketch through the same model and trains the sketch pass to reproduce how the photograph pass relates phrases to patches and patches to each other. Not one training label is drawn by hand. Sketch-Omni reaches 91.4 F1 on sketch-prompted localisation against 81.8 for the strongest general-purpose MLLM, with a lead that doubles when the photograph holds several instances of the queried category, and it is best in 39 of 40 evaluation cells across six tasks. Code, model, and a hand-annotated benchmark of 2754 sketch-to-photo links will be released.

1 INTRODUCTION

Perception has converged on a prompt. GroundingDINO (Liu et al., 2024) opened a detector's vocabulary with a text prompt. T-Rex2 (Jiang et al., 2024a) let a box or an exemplar stand in for the words. Rex-Omni (Jiang et al., 2026) folded detection, pointing, referring, and visual prompting into one MLLM that answers every prompt with coordinate tokens, and general-purpose MLLMs (Bai et al., 2025b;a; Yue et al., 2025; Wu et al., 2024) now ship grounding as standard. All of these models were built for photographs. None has been given a sketch. This paper brings sketch into the interface, in both directions (Fig. 1). **Sketch-Omni** works *on* sketches: the sketch is the image, and text prompts detect, point, resolve object and spatial referring expressions, and ground affordances in it. We call these five the *sketch-native* tasks. And it works *from* sketches: the sketch is the prompt, and given a drawing of one object and a photograph, the model localises that object in the photograph. To our knowledge it is the first perception MLLM to support both directions.

Sketch deserves this because it is not another kind of image. A photograph records what stood in front of a camera. A sketch records what a person has in mind: a product not yet built, a suspect never photographed. It is specific: a drawing shows the shape of one particular object, which is why fine-grained sketch retrieval has worked for a decade (Yu et al., 2016; Xu et al., 2023). And it is cheap: a few strokes, drawn in seconds, before any search begins. Put these together and a sketch can ask a question no other prompt can. A category name returns every instance of the category. A photo exemplar does the same, and it exists only if the object has already been photographed. A sketch, drawn from memory, asks for *that one* among several of its kind (Fig. 2). Chowdhury et al. (2023) first showed that a standalone detector can use a sketch this way. No perception MLLM can: its visual prompts are boxes and exemplars, and each stands for a category. We bring the query into the prompt interface as **instance-**

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.

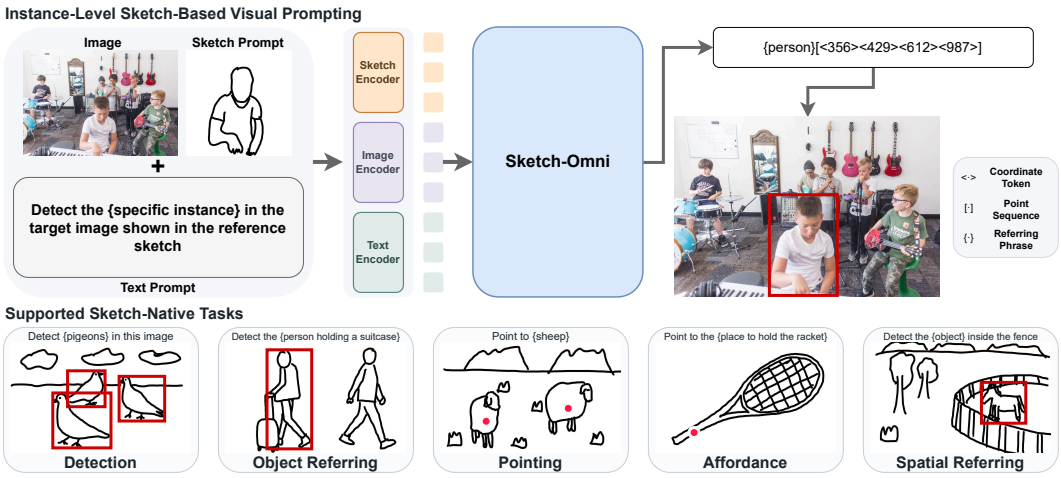

Figure 1: **Sketch-Omni.** Top, *from* a sketch: the sketch is the prompt, and a drawing of one object localises that object in a photograph. Bottom, *on* a sketch: the sketch is the image, and text prompts detect, point, resolve object and spatial referring expressions, and ground affordances in it. One model, one coordinate-token interface.

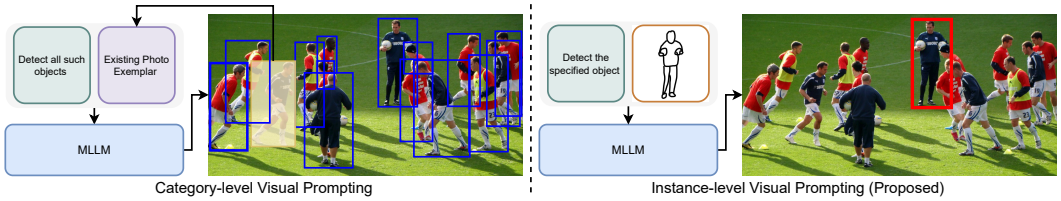

Figure 2: **Category-level versus instance-level visual prompting.** A photo exemplar returns every instance of its category (blue); a sketch returns only the one it depicts (red).

level sketch-based visual prompting (instance VP), and it is where a sketch model should prove itself: on photographs with several instances of the queried category, where a name no longer helps.

Two things stand in the way. The first is the model. A photo-trained MLLM grounds a phrase by relating its tokens to image patches, and it relates the patches of one object to each other. Both relations rest on attention tuned to the colour and texture of photographs. A sketch has neither: most of the canvas is blank, and every cue is a line. On a sketch, Rex-Omni's attention spreads over empty space (Fig. 5), and a general-purpose MLLM finds roughly the right region but not the strokes that bound it (Sec. 4.4). The second is data. Photographic grounding is trained on tens of millions of labelled photographs (Liu et al., 2024; Jiang et al., 2026). Sketch has nothing comparable: single-object drawings each paired with one photo (Sangkloy et al., 2016), scene sketches without boxes (Chowdhury et al., 2022), and instance links found by retrieval rather than by hand (Gao et al., 2020). Nobody will box-annotate a million drawings.

Sketch-Omni answers both with two ingredients. The first is **supervision built from photographs**. An edge detector (Xie & Tu, 2015) turns any labelled photograph into a sketch-like line drawing that keeps the photograph's boxes, phrases, and affordance labels; run over Objects365, RefCOCOg, and AGD20K, this supervises all five sketch-native tasks at no labelling cost. For instance VP we construct the pairs no dataset has: a Sketchy drawing whose paired photo is tiled into a canvas beside other photos of its category, and an edge-sketch crop of one annotated object in an Objects365 photograph, to be found again in that photograph. The second is **photo-guided relation alignment**, which teaches the model to read a sketch the way it already reads a photograph. Grounding rests on two relations: how much each phrase token attends to each image patch, and how the patches of one object relate to each other. Both are about structure, and structure is exactly what a photograph and its edge sketch share. So each labelled photograph passes through the frozen model and its edge sketch through the LoRA-adapted one, and the sketch pass is trained to reproduce the photograph pass's two relations at every layer and head. The features are left free to differ, since a sketch is not a photograph; only

the relations are matched. Not one training label is drawn by hand. The only hand annotation is for evaluation: 2754 sketch-to-photo instance links on FS-COCO, the first such benchmark we know of.

On instance VP, Sketch-Omni reaches 91.4 F1 at IoU 0.5 on FS-COCO, against 81.8 for the strongest general-purpose MLLM, Qwen3-VL-8B, given the same two images; its own base model, whose visual prompts must be boxes inside the target photograph, scores 4.0. The lead over Qwen3-VL-8B is 8.4 F1 when the photograph holds one instance of the queried category and 17.2 when it holds several (Sec. 4.3): the advantage sits exactly where a name fails. Across the six tasks Sketch-Omni is best in 39 of 40 table cells, and where general-purpose MLLMs collapse at tight boxes (Qwen3-VL-8B: 75.3 F1 at IoU 0.5 on free-hand detection, 23.0 at IoU 0.9) it holds 65.7. The alignment improves every one of the 14 ablation cells, by up to 4.3 F1 on referring, the task that depends most on a phrase selecting its region (Table 1, App. F).

What this opens. Sketch research has built one model per task: a recogniser, a retriever, a segmenter. Sketch-Omni replaces the model with a prompt. Any task the grounding literature poses to a photograph can now be posed to a drawing, in the same words; this paper poses six. For grounding research, sketch offers three things: a visual prompt that names an instance without a photograph of it; a modality added to a perception MLLM without a single new label; and a test of how much a model's grounding leans on appearance, since a sketch is a photograph with its appearance removed, and the IoU 0.9 column is the score.

2 Related Work

Promptable perception. GroundingDINO (Liu et al., 2024) made a DETR-style detector (Zhang et al., 2023) open-set by conditioning it on a text prompt, T-Rex2 (Jiang et al., 2024a) added visual prompts beside the text, and ChatRex (Jiang et al., 2024b) carried prompted perception into a multimodal LLM. In parallel, Pix2Seq (Chen et al., 2022) cast detection as language modelling over coordinate tokens. Autoregressive coordinate prediction was subsequently adopted by grounding MLLMs (Peng et al., 2024; Chen et al., 2023; You et al., 2024; Zhan et al., 2024; Wang et al., 2024; Xiao et al., 2024) and is now built into general-purpose MLLMs (Bai et al., 2025b;a; Wu et al., 2024; Deitke et al., 2025; Yue et al., 2025; Guo et al., 2025). Rex-Omni (Jiang et al., 2026) closes the loop, unifying detection, pointing, referring, OCR, keypoints, visual prompting, and GUI grounding under one coordinate-token vocabulary. Where language cannot name a target, they take clicks, boxes, masks, and regions of interest as prompts (Kirillov et al., 2023; Zou et al., 2023; Zhang et al., 2024; Lin et al., 2025; Yang et al., 2023), each placed on the image being perceived, so the object is already in view, or taken from another photograph as a category exemplar (Jiang et al., 2024a). The whole line assumes captured images to perceive and photographs to prompt with. Sketch-Omni adds a modality to this line, not a task.

Sketch as a query. Sketch research has argued for the drawing as a query since its datasets were built (Eitz et al., 2012; Sangkloy et al., 2016; Ha & Eck, 2018; Gao et al., 2020; Chowdhury et al., 2022), chiefly through sketch-based image retrieval (Xu et al., 2023; Dey et al., 2019; Lin et al., 2023), and fine-grained SBIR (Yu et al., 2016; Sain et al., 2020) established that a drawing identifies an instance rather than a category. Retrieval returns whole images, though. Tripathi et al. (2020; 2024) moved the query inside the image, localising sketched objects in natural images, and Chowdhury et al. (2023) build a zero-shot detector on a CLIP-initialised (Radford et al., 2021) SBIR backbone that takes the sketch as an instance-aware query. O3SLM (Gupta et al., 2026) places the sketch query in an LVLM trained on image–sketch–instruction triplets for localisation, counting, retrieval, and question answering. Its sketch stands for a category, since a drawing is paired with any image that holds the drawn class and localisation returns every instance of it, the behaviour of the photo exemplar in Figure 2, and it never perceives within a sketch. The instance-level query therefore exists only in standalone detectors that take no language and serve one task.

Perceiving within sketches. Understanding what a sketch contains has been left to single-task models. Scene sketches have been segmented semantically, recently with an open vocabulary (Bourouis et al., 2024), and by instance, where InkLayer (Tang et al., 2025) fine-tunes GroundingDINO on scene sketches class-agnostically, collapsing every label to *object*, so the detector survives the domain gap but the prompt does not. SketchFusion (Koley et al., 2025) fuses foundation-model features into a universal sketch encoder and StructXLIP (Ruan et al., 2026) aligns CLIP embeddings against

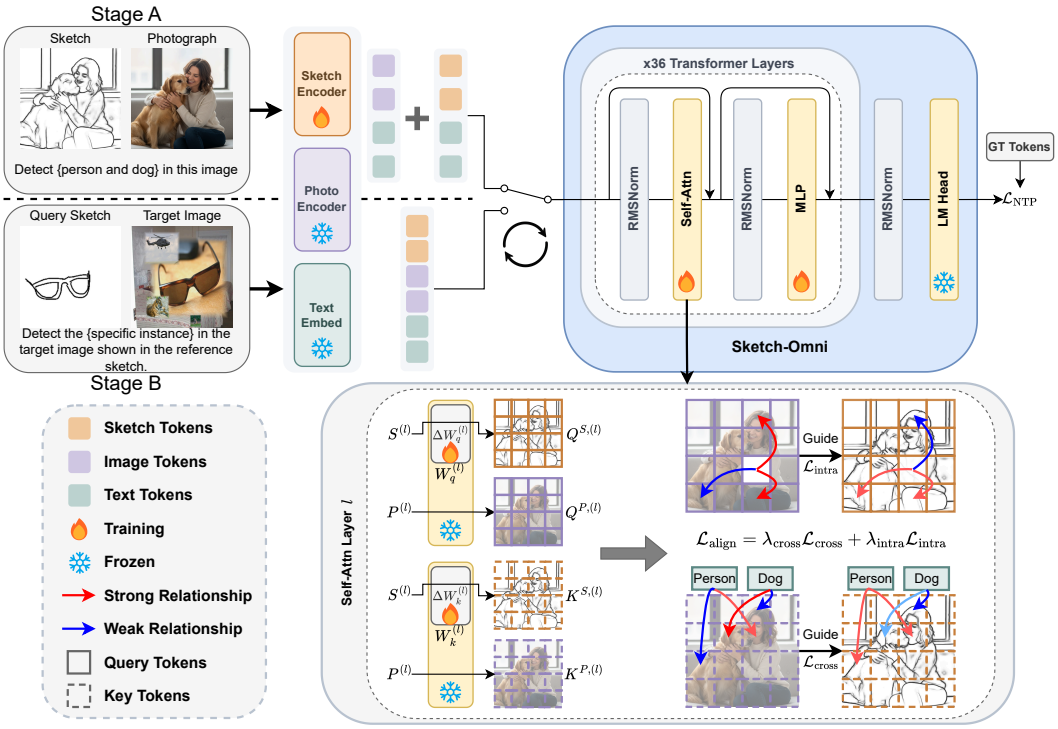

Figure 3: **Sketch-Omni overview.** Left, both stages share one LoRA-adapted backbone under a single optimiser. Right, photo-guided relation alignment matches the sketch path's token relations to the frozen photo path's via $\mathcal{L}_{\text{cross}}$ and $\mathcal{L}_{\text{intra}}$, with solid grids denoting queries and dashed grids keys. $S^{(l)}$ and $P^{(l)}$ are the two paths' hidden states at layer l, and $\Delta W_q^{(l)}$, $\Delta W_k^{(l)}$ the LoRA updates to the frozen projections $W_q^{(l)}$, $W_k^{(l)}$.

structural cues, but none is prompted beyond a category list, and none points, refers, or grounds affordances on a drawing. Sketch-Omni is where the two lines meet.

3 Sketch-Omni

Sketch-Omni adds a modality to Rex-Omni (Jiang et al., 2026) and leaves its interface unchanged. Stage A (Sec. 3.1) transfers the base model's grounding to sketch input through photo-guided relation alignment while training the five sketch-native tasks, and Stage B (Sec. 3.2) builds instance VP on that shared reading from constructed supervision.

The interface casts perception as next-token prediction. A textual prompt T and the visual inputs form one context, and coordinates are discretised into 1000 bins, each a dedicated coordinate token (App. A). *Sketch-native perception* makes the sketch the image input, and the prompt T lists the phrases to be grounded. A LoRA (Hu et al., 2022) update $\Delta\theta$ on the frozen pretrained weights θ_0 defines the *sketch path* and θ_0 alone the *photo path*, written S and P for the input each runs on. Both training stages fit $\Delta\theta$ with a next-token prediction loss $\mathcal{L}_{\text{NTP}}$ on the target sequence, and they interleave under a single optimiser, alternating at a fixed step interval (Sec. 4.2).

3.1 Stage A: Photo-guided relation alignment

Stage A trains the five sketch-native tasks, detection, pointing, object referring, spatial referring, and affordance grounding, from two sources. A labelled photograph yields its pixel-aligned edge sketch through a learned detector (Xie & Tu, 2015), so the sketch trains under supervision carried over from the photograph. Since a sketch cannot show colour or texture, a lexicon filter (Ruan et al., 2026) strips appearance terms from each phrase; only the phrases it alters are re-grounded on the photograph by the base model, whose answers become their labels, and every other label stays as annotated. What stays invariant across the two renderings is the scene structure, which patches

belong to one object and which region a phrase selects, so the alignment supervises these relations and leaves the features free: it never asks the sketch path to look like the photo path, only to relate its tokens as the photo path relates its own, and the sketch path still encodes strokes.

Which relations to align. Grounding rests on two kinds of relation. A phrase has to select its region, and attention already computes this as the cross-modal query-to-key relation, which $\mathcal{L}_{\text{cross}}$ aligns across the two paths. The parts of one object have to hold together, and for this we borrow the intra-modal query-to-query relation underlying DeCLIP's context features (Wang et al., 2025), which $\mathcal{L}_{\text{intra}}$ aligns in the same way. Both losses compare corresponding token positions between the two readings of one scene, not samples in a batch as in relation-based distillation (Park et al., 2019; Tung & Mori, 2019).

How the relations are matched. The photo path runs without gradients, so the alignment adds one forward pass per Stage-A step (Fig. 3). The sketch keeps the photograph's prompt, target sequence, and patch grid, so a relation in one path has an exact counterpart in the other.

At every language-model decoder layer l, the sketch path computes $Q^{S,(l)}$ and $K^{S,(l)}$ through the adapted projections and the photo path computes $Q^{P,(l)}$ and $K^{P,(l)}$ through the frozen ones (Fig. 3). We align every head of every layer, where each key head is shared by a group of query heads. Cosine similarities are computed after Qwen2.5-VL's rotary position embeddings (Bai et al., 2025b) are applied. Let $\mathcal{V}$ hold the positions of the image tokens, $\mathcal{F} \subseteq \mathcal{V}$ those whose patch overlaps any annotated box, and $\mathcal{T}$ the positions of the phrases in the prompt, all taken on the input side of the teacher-forced sequence. $\mathcal{L}_{\text{cross}}$ keeps every visual position as a key, which lets it penalise phrase similarity that falls on blank regions. $\mathcal{L}_{\text{intra}}$ restricts its queries to $\mathcal{F}$, which keeps the grouping signal on foreground object patches. For simplicity, q_i^S and k_j^S denote path S's query and key at positions i and j, with the layer and head indices suppressed. With $\Phi(u, v) = \langle u, v\rangle/(\|u\|\|v\|)$, $\mathcal{D}$ the decoder layers, and H the number of heads,

$$\mathcal{L}_{\text{cross}} = \frac{1}{|\mathcal{D}|}\sum_{l\in\mathcal{D}}\sum_{h=1}^{H}\sqrt{\frac{1}{|\mathcal{T}||\mathcal{V}|}\sum_{i\in\mathcal{T},\, j\in\mathcal{V}}\big(\Phi(q_i^S, k_j^S) - \Phi(q_i^P, k_j^P)\big)^2}, \tag{1}$$

$$\mathcal{L}_{\text{intra}} = \frac{1}{|\mathcal{D}|}\sum_{l\in\mathcal{D}}\sum_{h=1}^{H}\sqrt{\frac{1}{|\mathcal{F}|^2}\sum_{i,\, j\in\mathcal{F}}\big(\Phi(q_i^S, q_j^S) - \Phi(q_i^P, q_j^P)\big)^2}. \tag{2}$$

With $\mathcal{L}_{\text{align}} = \lambda_{\text{cross}}\,\mathcal{L}_{\text{cross}} + \lambda_{\text{intra}}\,\mathcal{L}_{\text{intra}}$, Stage A optimises

$$\mathcal{L}_A \;=\; \mathcal{L}_{\text{NTP}} + \mathcal{L}_{\text{align}} \tag{3}$$

jointly over the five sketch-native tasks. The two weights are given in Sec. 4.2.

3.2 Stage B: instance-level visual prompting

Stage B places two images in the context, a target photograph first and a query sketch second, followed by a prompt naming the sketch as the query (App. A). The answer is a match between the sketch's tokens and the photograph's, which is what the shared reading built in Stage A makes possible. No dataset annotates the supervision this needs, a free-hand drawing tied to the one instance it depicts inside a larger scene, so we build it in two ways.

The first construction uses sketch–photo pair datasets, in which a free-hand drawing is paired with a photograph of a single object. We tile the paired photograph, together with crops of the same category, onto a low-clutter background, so the drawing designates exactly one instance among them, and the category name alone cannot identify the target. The second construction supplies a real scene, with its clutter and occlusion, by deriving pairs from ordinary detection annotations. Given an annotated object, we crop its region from the photograph's edge sketch and use the crop as the query, asking the model to localise that instance in the untouched photograph. One construction trains tolerance to variation in drawing style, the other precise shape matching in real scenes. Full construction details are given in App. B.

When an input holds both a photograph and a sketch, in Stage B and at inference, the vision encoder runs once per image in the interface's fixed image order, with the LoRA update disabled for the photograph and applied for the sketch; the language model applies it over the whole sequence.



Figure 4: **Qualitative instance-level sketch VP.** Each row shows a free-hand query sketch, the target photograph with the ground-truth box, and one panel per model.

Stage B trains with next-token prediction alone, interleaved with Stage A, so the alignment keeps acting on the sketch path while the prompting capability forms on top of it. Alongside the two-image samples, detection and pointing train in this stage on multi-object canvases composited from single-object free-hand drawings, where prompts naming absent categories are answered with an explicit *None*. A Group Relative Policy Optimisation (GRPO) (Shao et al., 2024) pass with localisation rewards then refines the model (App. G), and training sources for both stages are given in Sec. 4.1. If shape is what lets a query pick out its instance, the advantage over a competent baseline should widen once same-category distractors are present (Sec. 4.3).

4 Experiments

4.1 Datasets

Stage A, sketch-native perception. Each sketch-native task takes photographs from a photo-grounding dataset and feeds the matching HED edge sketch (Xie & Tu, 2015) as the input. Objects365 (Shao et al., 2019) supervises detection and pointing, and box centres serve as the pointing targets. RefCOCOg (Mao et al., 2016) supervises object referring and spatial referring, which name an instance by what it is or by where it stands, and AGD20K (Luo et al., 2022) affordance grounding.

Stage B, instance-level sketch-based visual prompting. Stage B trains instance-level visual prompting on **Sketchy** (Sangkloy et al., 2016), which pairs free-hand sketches with single-object photos. One positive crop is tiled alongside same-category distractor crops onto a low-clutter Objects365 background. No Sketchy sketch of a SketchyCOCO foreground category enters this training, and any training sketch that also appears as a SketchyCOCO query is removed. The second construction draws on Objects365 (Shao et al., 2019), where the query is an annotated object's edge sketch cropped at its box and the target is the source photograph.

Evaluation. We evaluate against base Rex-Omni (Jiang et al., 2026) and baselines appropriate to each task. Detection, referring, and pointing are compared against the general perception MLLMs Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), MiMo-VL (Yue et al., 2025), and DeepSeek-VL2 (Wu et al., 2024). Instance VP is compared against the same four, where each model receives the query sketch and the target photograph as two images. Since these four MLLMs are not specifically trained on sketches, we also compare against a pipeline that pairs a detector with a sketch retrieval model: GroundingDINO (Liu et al., 2024) proposes boxes in the target photograph under a fixed vocabulary prompt shared by every query, and ZSE-SBIR (Lin et al., 2023) ranks them against the query sketch. Affordance is compared against the pointing specialists Molmo (Deitke et al., 2025), RoboPoint (Yuan et al., 2024), and RoboRefer (Zhou et al., 2025). Referring and affordance

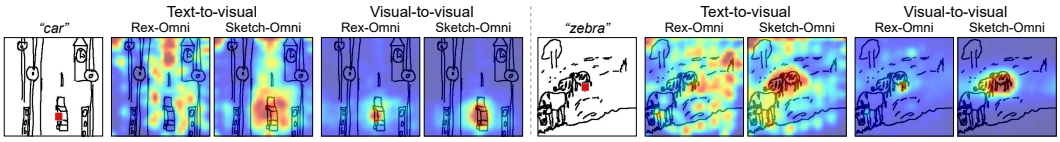

Figure 5: **Cross-modal and intra-modal relations on free-hand sketches.** Cross-modal maps show phrase-query similarity to every visual key, and intra-modal maps the red-marked patch's query against every visual query. All maps are from layer 22.

Table 1: **Instance-level sketch VP.** Each cell is SketchyCOCO / FS-COCO, and the lower block ablates the alignment. **Bold** means best and underline means second best.

Method	F1@0.3	F1@0.5	F1@0.7
Qwen2.5-VL-3B	59.3/55.8	53.6/51.1	48.8/47.5
Qwen3-VL-8B	68.3/84.5	66.5/81.8	62.5/76.9
MiMo-VL-7B	63.6/48.2	61.2/45.5	57.2/40.1
DeepSeek-VL2-S	53.0/43.1	50.3/36.0	46.5/26.0
Rex-Omni	18.1/5.7	13.0/4.0	4.7/1.4
G-DINO + ZSE-SBIR	69.5/55.6	63.3/51.1	57.9/47.1
Sketch-Omni (full)	**74.0/93.6**	**71.0/91.4**	**67.7/87.1**
w/o align	73.1/91.2	69.8/88.6	66.7/84.2

Table 2: **Stratified instance VP.** Each cell is SketchyCOCO / FS-COCO on the Easy and Hard subsets of Sec. 4.3, and the Δ rows report Sketch-Omni minus Qwen3-VL-8B.

Method	Subset	F1@0.3	F1@0.5	F1@0.7
Qwen3-VL-8B	Easy	91.0/86.9	**90.2**/84.3	**86.7**/79.5
	Hard	50.5/66.2	47.8/63.1	43.4/56.8
Sketch-Omni	Easy	**91.7/94.8**	89.5/**92.7**	86.1/**88.3**
	Hard	**55.3/83.3**	**51.4/80.3**	**48.2/77.0**
Δ	Easy	+0.7/+7.9	−0.7/+8.4	−0.6/+8.8
	Hard	+4.8/+17.1	+3.6/+17.2	+4.8/+20.2

are evaluated on HED sketches. These are pixel-aligned with the source photographs, so we reuse the photographs' ground truth. Detection and pointing are evaluated on free-hand sketches from FS-COCO (Chowdhury et al., 2022) and on HED sketches rendered from COCO (Lin et al., 2014) under Rex-Omni's protocol.

Free-hand evaluation has no such transfer and requires new labels. SketchyCOCO (Gao et al., 2020) pairs real photographs with composited scene sketches. However, three properties limit what a margin on it can show: (1) Its queries are linked to their instances by a retrieval model (Sangkloy et al., 2016), so the links are only as accurate as that model. (2) The queries come from a small pool, so the same drawing recurs across different photographs. (3) The queried instance is typically the photograph's dominant object, covering 17.1% of it at the median against 8.8% on FS-COCO, so localising the most salient object succeeds without attending to the query. FS-COCO provides what SketchyCOCO cannot, genuine free-hand scene sketches drawn from memory after viewing the paired photograph, but it provides neither object boxes nor instance links. We therefore annotate its test split with boxes on the sketch as ground truth for detection and pointing, and with sketch-to-photo instance links for VP, where each query is the region of the scene sketch that the annotated box covers (App. C). The evaluation comprises 1340 SketchyCOCO and 2754 FS-COCO sketch queries, with qualitative results in Figure 4.

Metrics. Given predicted and ground-truth boxes, we match predictions to ground truth one-to-one and greedily by Intersection-over-Union (IoU). Recall and precision are computed over all matches in an evaluation set, and F1 is their harmonic mean. Detection and referring report IoU 0.5, IoU 0.9, and the mean across thresholds from 0.5 to 0.95 with step size 0.05, shown in the *avg* columns. Instance VP reports IoU 0.3, 0.5, and 0.7. Pointing and affordance test containment instead. For pointing, a prediction counts as correct when it falls inside the target's segmentation mask, or its box where no mask exists. For affordance, it must fall inside the ground-truth region obtained by thresholding the heatmap at half its peak (App. D).

4.2 Implementation Details

We initialise θ_0 from Rex-Omni (Jiang et al., 2026), a Qwen2.5-VL-3B (Bai et al., 2025b) backbone, and train rank-64 LoRA adapters $\Delta\theta$ (Hu et al., 2022) on all linear layers of every language-model block and on the query, key, value, and output projections of the vision encoder, with θ_0 frozen. The two stages alternate every 50 optimiser steps at a Stage-A to Stage-B ratio of 2:3, and the Stage-A objective (Eq. 3) uses $\lambda_{\text{cross}}=0.5$ and $\lambda_{\text{intra}}=0.25$. Decoding at inference is greedy. Every configuration, ablations included, is refined by the same final GRPO pass following Rex-Omni. Optimiser and GRPO settings are given in App. G.

Table 3: **Sketch-input detection and pointing.** Each cell is FS-COCO / COCO.

Method	Detection									Pointing
	R@0.5	P@0.5	F1@0.5	R@0.9	P@0.9	F1@0.9	R@avg	P@avg	F1@avg	F1@Point
Qwen2.5-VL-3B	48.8/24.2	69.7/41.4	57.4/30.6	13.3/6.5	19.1/13.2	15.7/8.7	30.6/14.8	43.7/27.4	36.0/19.2	72.2/36.3
Qwen3-VL-8B	70.2/34.3	81.2/55.6	75.3/42.4	21.4/9.9	24.8/17.7	23.0/12.7	47.5/22.0	55.0/36.7	51.0/27.5	81.7/**47.9**
MiMo-VL-7B	39.6/19.1	43.5/30.2	41.5/23.4	1.9/2.5	2.1/5.6	2.0/3.4	16.6/8.9	18.3/16.3	17.4/11.5	61.8/33.8
DeepSeek-VL2-S	57.8/31.5	62.3/50.0	60.0/38.6	11.6/9.3	12.5/15.8	12.0/11.7	33.8/19.9	36.5/32.3	35.1/24.7	-
Rex-Omni	74.7/24.3	86.0/55.1	80.0/33.7	52.4/7.2	60.4/18.9	56.1/10.4	63.5/15.5	73.1/37.5	68.0/22.0	84.9/37.9
Sketch-Omni	**81.9/34.5**	**92.1/79.2**	**86.7/48.1**	**62.1/15.8**	**69.8/37.3**	**65.7/22.2**	**72.4/25.7**	**81.4/59.7**	**76.6/36.0**	**89.9**/44.5

4.3 Instance-Level Sketch-Based Visual Prompting

Sketch-Omni outperforms every baseline at all three IoU thresholds (Table 1). Base Rex-Omni fails at this task: its visual prompting takes reference boxes inside the target image and returns every similar object, so it has no instance-level query mode and seldom answers the two-image prompt. Sketch-Omni leads the strongest MLLM by 4.5 and 9.6 F1 at IoU 0.5 on SketchyCOCO and FS-COCO, and the assembled pipeline by 7.7 and 40.3, since its retrieval stage ranks proposals by category and cannot pick one instance out of several. The lower block ablates the alignment. The *w/o align* row is the same recipe trained without $\mathcal{L}_{\text{align}}$, and the constructed supervision alone establishes the capability, outperforming every baseline in every cell. The alignment provides a consistent further improvement of 0.9 to 2.9 points in all six cells. Figure 5 shows how base Rex-Omni and Sketch-Omni differ on a sketch-native detection prompt: cross-modal similarity from the phrase spreads over blank regions under Rex-Omni and concentrates on the queried object under Sketch-Omni, and intra-modal similarity propagates from a token across its instance only under Sketch-Omni. Figure 4 shows the failure modes: a baseline that misses returns a same-category distractor, another salient object, or nothing, where Sketch-Omni localises the instance the sketch depicts.

Resolving same-category distractors. Instance VP is defined by discrimination among same-category distractors, which single-instance scenes never exercise. We therefore split both benchmarks by the number of annotated instances of the queried category in the photograph, *Easy* when it holds exactly one and *Hard* when it holds two or more. Counts and all MLLM baselines are in App. E. Table 2 shows that on FS-COCO-Hard, Sketch-Omni reaches 80.3 F1 at IoU 0.5 against 63.1 for the strongest baseline, 17.2 above it. SketchyCOCO shows the same ordering with a smaller margin, since its same-category instances differ little from one another and leave less for a query to separate. The two models are within one point of each other on SketchyCOCO-Easy. On FS-COCO, whose queries are from the target's own strokes, Sketch-Omni leads by 7.9 to 8.8 F1 even on Easy, and the lead grows with the IoU threshold as a faithful query improves the box itself.

Table 4: **Sketch-input referring on RefCOCOg.** Each cell is val / test. Object and spatial referring are told apart by the presence of a spatial term in the expression.

Method	Object referring			Spatial referring		
	F1@0.5	F1@0.9	F1@avg	F1@0.5	F1@0.9	F1@avg
Qwen2.5-VL-3B	51.5/50.8	15.8/16.7	33.7/33.1	50.3/50.0	13.6/12.5	30.2/30.2
Qwen3-VL-8B	62.2/62.4	23.7/25.5	43.8/44.3	61.7/62.5	21.1/21.9	42.3/42.8
MiMo-VL-7B	43.5/45.7	2.3/3.4	20.0/22.4	41.7/42.7	1.8/2.4	18.4/18.6
DeepSeek-VL2-S	69.1/66.1	25.4/25.9	48.2/46.5	67.6/65.1	21.1/20.2	43.9/43.0
Rex-Omni	52.0/51.2	20.2/20.3	36.3/35.7	54.2/54.4	19.4/19.8	37.1/37.5
Sketch-Omni	**75.0/74.6**	**43.1/42.7**	**59.6/59.6**	**77.6/76.7**	**43.3/43.1**	**59.9/60.7**

Table 5: **Affordance grounding on AGD20K.** Point-in-region F1 on the Seen and Unseen splits, against pointing specialists.

Method	Seen	Unseen
Molmo-7B-D	53.6	56.8
Molmo-72B	56.8	58.6
RoboPoint	13.8	13.9
RoboRefer-2B	40.6	43.7
Rex-Omni	51.3	61.0
Sketch-Omni	**66.2**	**74.3**

4.4 Sketch-Native Tasks

Sketch-Omni achieves the best recall, precision, and F1 in every detection column of Table 3, with qualitative results in Figure 6. On free-hand sketches, Sketch-Omni attains its largest margin at F1@0.9, leading base Rex-Omni by 9.6 and the best general-purpose MLLM by 42.7. Sketch-Omni retains 62.1 of its 81.9 recall as the threshold tightens from 0.5 to 0.9, while Qwen3-VL drops from 70.2 to 21.4, the signature of grounding tuned to appearance: the region is found and the strokes are not. On edge sketches, Sketch-Omni leads the best baseline by 23.6 in precision at IoU 0.5. On pointing, Sketch-Omni outperforms every model on free-hand sketches.

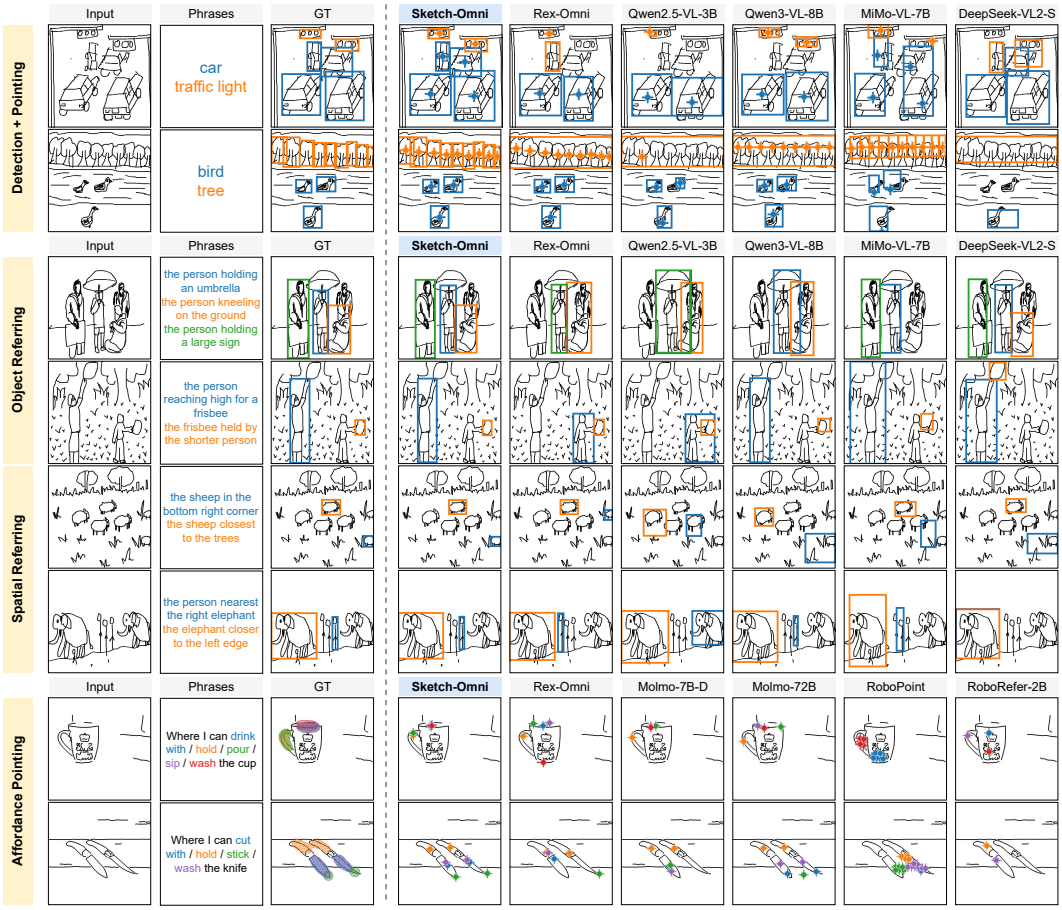

Figure 6: **Qualitative results on sketch input.** Two examples each of detection and pointing, object referring, spatial referring, and affordance grounding, with box and point colours matching the phrase colours. The referring expressions and their ground-truth boxes are newly annotated for these examples, and DeepSeek-VL2-S outputs boxes only.

On object and spatial referring (Table 4), Sketch-Omni achieves the best result in every cell, and attains its largest margin at IoU 0.9, where object F1 reaches 43.1/42.7 against 25.4/25.9 for the strongest baseline. Every general-purpose MLLM loses average F1 on spatial expressions relative to its own object subset, up to 4.3 points for the strongest, while Sketch-Omni widens its margin from 11.4/13.1 to 16.0/17.7 and resolves spatial expressions better than object ones, 77.6 against 75.0 F1 at IoU 0.5 on val, which is what a drawing that keeps layout and drops appearance should do.

On AGD20K affordance grounding (Table 5), the 3B Sketch-Omni outperforms every pointing specialist on both splits, improving base Rex-Omni by 14.9 and 13.3. The alignment accounts for 0.4 to 4.3 F1 of the gain across the five tasks (App. F). Across Tables 1, 3, 4, and 5, Sketch-Omni is best in 39 of 40 cells.

5 Conclusion

Sketch-Omni brings promptable perception to sketch, on the drawing and from it. One diagnosis, that a sketch keeps structure and drops appearance, gives the recipe: construct the supervision, and align relations rather than features, with no new training label. The evidence concentrates where a name fails, among several instances of one category, and there Sketch-Omni leads the strongest baseline by 3.6 F1 on SketchyCOCO and 17.2 on FS-COCO at IoU 0.5. For sketch research, every perception task is now a prompt away; for grounding, sketch is a prompt that carries identity, a modality that costs no labels, and a test of what grounding leans on.

Reproducibility statement

The token interface, the construction of all synthetic training data, the benchmark annotation protocol, and the GRPO configuration are specified in App. A–G, with training and evaluation settings in Secs. 4.2 and 4.1. Our code and annotated benchmark will be released.

AI use statement

In this work, we used generative AI tools to edit the text of this paper for readability, specifically for rephrasing sentences, checking grammar, and improving the flow of the writing. Generative AI tools were not used to propose the research idea, to design the method or the experiments, or to produce any experimental result. All experiments were run and analysed by the authors. We have reviewed all AI-assisted work and take full responsibility for the content of this paper.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Ahmed Bourouis, Judith Ellen Fan, and Yulia Gryaditskaya. Open vocabulary semantic scene sketch understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *International Conference on Learning Representations (ICLR)*, 2022.

Pinaki Nath Chowdhury, Aneeshan Sain, Ayan Kumar Bhunia, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. FS-COCO: Towards understanding of freehand sketches of common objects in context. In *European Conference on Computer Vision (ECCV)*, 2022.

Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. What can human sketches do for object detection? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Sounak Dey, Pau Riba, Anjan Dutta, Josep Lladós, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *Acm Transactions on Graphics (TOG)*, 2012.

Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. SketchyCOCO: Image generation from freehand scene sketches. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1.5-VL technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Rishi Gupta, Mukilan Karuppasamy, Shyam Marjit, Aditay Tripathi, and Anirban Chakraborty. O3SLM: Open weight, open data, and open vocabulary sketch-language model. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

David Ha and Douglas Eck. A neural representation of sketch drawings. In *International Conference on Learning Representations (ICLR)*, 2018.

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 (ICLR)*, 2022.

Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-Rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision (ECCV)*, 2024a.

Qing Jiang, Gen Luo, Yuqin Yang, Yuda Xiong, Yihao Chen, Zhaoyang Zeng, Tianhe Ren, and Lei Zhang. ChatRex: Taming multimodal LLM for joint perception and understanding. *arXiv preprint arXiv:2411.18363*, 2024b.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Subhadeep Koley, Tapas Kumar Dutta, Aneeshan Sain, Pinaki Nath Chowdhury, Ayan Kumar Bhunia, and Yi-Zhe Song. SketchFusion: Learning universal sketch features through fusing foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Fengyin Lin, Mingkang Li, Da Li, Timothy Hospedales, Yi-Zhe Song, and Yonggang Qi. Zero-shot everything sketch-based image retrieval, and in explainable style. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable MLLMs to comprehend what you want. In *International Conference on Learning Representations (ICLR)*, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, 2024.

Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. In *International Conference on Learning Representations (ICLR)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Zanxi Ruan, Songqun Gao, Qiuyu Kong, Yiming Wang, and Marco Cristani. StructXLIP: Enhancing vision-language models with multimodal structural cues. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. In *British Machine Vision Conference (BMVC)*, 2020.

Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *Acm Transactions on Graphics (TOG)*, 2016.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Mia Tang, Yael Vinker, Chuan Yan, Lvmin Zhang, and Maneesh Agrawala. Instance segmentation of scene sketches using natural image priors. In *SIGGRAPH Conference Papers '25*, 2025.

Aditay Tripathi, Rajath R. Dani, Anand Mishra, and Anirban Chakraborty. Sketch-guided object localization in natural images. In *European Conference on Computer Vision (ECCV)*, 2020.

Aditay Tripathi, Anand Mishra, and Anirban Chakraborty. Query-guided attention in vision transformers for localizing objects using a single sketch. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. DeCLIP: Decoupled learning for open-vocabulary dense perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. CogVLM: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Peng Xu, Timothy M. Hospedales, Qiyue Yin, Yi-Zhe Song, Tao Xiang, and Liang Wang. Deep learning for free-hand sketch: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 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 (ICLR)*, 2024.

Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen-Change Loy. Sketch me that shoe. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. RoboPoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. MiMo-VL technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *European Conference on Computer Vision (ECCV)*, 2024.

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 (ICLR)*, 2023.

Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, and Yinfei Yang. Ferret-v2: An improved baseline for referring and grounding with large language models. In *Conference on Language Modeling (COLM)*, 2024.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. RoboRefer: Towards spatial referring with reasoning in vision-language models for robotics. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

APPENDIX

A TOKEN INTERFACE OF THE BASE MODEL

Rex-Omni (Jiang et al., 2026) casts every perception task as conditional sequence generation over one shared vocabulary, so detection, pointing, referring, and visual prompting differ only in their prompt and output content.

Input grammar. A visual input is encoded by the vision encoder into N_v patch tokens under Qwen2.5-VL's dynamic-resolution convention, so N_v varies with the input's size and aspect ratio. The patch tokens are wrapped in the image delimiters $\langle\text{im}\rangle$ and $\langle/\text{im}\rangle$, then concatenated with the N_t tokens of the textual prompt T,

$$\mathbf{x} \;=\; \big[\,\langle\text{im}\rangle,\, v_1,\,\ldots,\,v_{N_v},\,\langle/\text{im}\rangle,\, w_1,\,\ldots,\,w_{N_t}\,\big]. \tag{4}$$

For sketch-native tasks T carries phrases $p_1, \ldots, p_M$, as in "Detect $p_1, \ldots, p_M$ in the sketch.", where M is the number of phrases and a phrase may be a category name, a referring expression, or an affordance verb. Sketch VP uses a fixed prompt, since the query is carried by the sketch span rather than by words. The position sets that Sec. 3 pairs index this sequence: $\mathcal{V}$ holds the positions of $v_1, \ldots, v_{N_v}$, $\mathcal{T}$ those of the tokens of $p_1, \ldots, p_M$ within $w_1, \ldots, w_{N_t}$, and $\mathcal{F} \subseteq \mathcal{V}$ those whose patch overlaps an annotated box.

Output grammar. Coordinates are normalised to the image and quantised into 1000 bins, each a dedicated *coordinate token* in the vocabulary. Every task answers with one structured token sequence, built from a descriptive phrase, coordinate tokens, and special tokens that demarcate the two,

$$\underbrace{\langle\text{ref}\rangle\ \textsc{phrase}\ \langle/\text{ref}\rangle}_{\text{phrase}}\ \underbrace{\langle\text{box}\rangle\ \textsc{coords}\ \langle/\text{box}\rangle}_{\text{geometry}}, \tag{5}$$

where PHRASE names the object the coordinates stand for, and COORDS lists them, $[x_0, y_0, x_1, y_1]$ per box and $[x_0, y_0]$ per point. Several instances of one phrase share a single geometry segment, comma-separated and ordered by increasing x_0, and commas likewise join the blocks of different phrases. A phrase with no match in the input carries a literal *None* inside the geometry segment.

Visual prompting in the base model. Rex-Omni prompts visually with boxes rather than with a second image. One or more reference boxes inside the target image are quantised into the same coordinate tokens as an output box and passed in the prompt as a list under a single object key, and the released template reads "Given reference boxes {"object_1": ["$c_1c_2c_3c_4$", ...]} indicating one or more objects, find all similar objects in the image and output their bounding boxes". The answer lists every box of the indicated category. The query therefore names a category by example and lives in the image it queries, which is why it has no way to ask for one instance among several of that category (Sec. 4.3).

Sketch-VP extension. The sketch-VP input extends Eq. 4 to two delimited visual spans, the target photograph first and the query sketch second, with the prompt specifying the query sketch,

$$\mathbf{x}_{\text{vp}} \;=\; \big[\,\langle\text{im}\rangle\, v_1^P,\,\ldots,\,v_{N_P}^P,\,\langle/\text{im}\rangle,\,\langle\text{im}\rangle\, v_1^S,\,\ldots,\,v_{N_S}^S,\,\langle/\text{im}\rangle,\, w_1,\,\ldots,\,w_{N_t}\,\big], \tag{6}$$

where $v_1^P, \ldots, v_{N_P}^P$ are the patch tokens of the target photograph and $v_1^S, \ldots, v_{N_S}^S$ those of the query sketch. The output grammar is Eq. 5 unchanged, with the queried instance's category name filling the reference segment and the answer box expressed in the target photograph's coordinate frame. Sketch-Omni therefore inherits Rex-Omni's decoding, parsing, and evaluation stack verbatim.

B TRAINING DATA CONSTRUCTION

Composited sketch scenes for Stage B detection and pointing. The scenes composite single-object drawings from three sources, each contributing a different kind. QuickDraw (Ha & Eck, 2018) holds 345 categories of doodles drawn under a twenty-second limit, TU-Berlin (Eitz et al., 2012) 250 categories drawn under a much looser thirty-minute limit, and Sketchy (Sangkloy et al., 2016) 125 categories drawn from memory after a brief view of a photograph, so mixing them exposes

Stage B to the range of abstraction a free-hand query can take. All three store strokes as vectors. Their class names are unified into a self-derived vocabulary of 415 categories, the canonicalised union of the three: names are lowercased, de-pluralised, and passed through a synonym table that merges variants such as *aeroplane* and *airplane*. For each instance, a source dataset is first sampled uniformly, then a category within that source, then a drawing, so the mix stays near one third per source and QuickDraw's far larger pool cannot dominate. Each scene holds 2 to 12 instances from 2 to 6 categories on a canvas whose long side varies from 768 to 1152 pixels (Fig. 7). Every drawing is rendered from its stroke data directly at its final canvas size rather than rescaled, so stroke width stays constant across object sizes. A placement is rejected when the new box overlaps a placed one above 0.5 IoU, or when either box contains more than 0.8 of the other, so no object ends up hidden inside another. Prompts list the present categories together with sampled absent ones, which the target answers with *None*. Scenes are generated procedurally during training rather than stored.

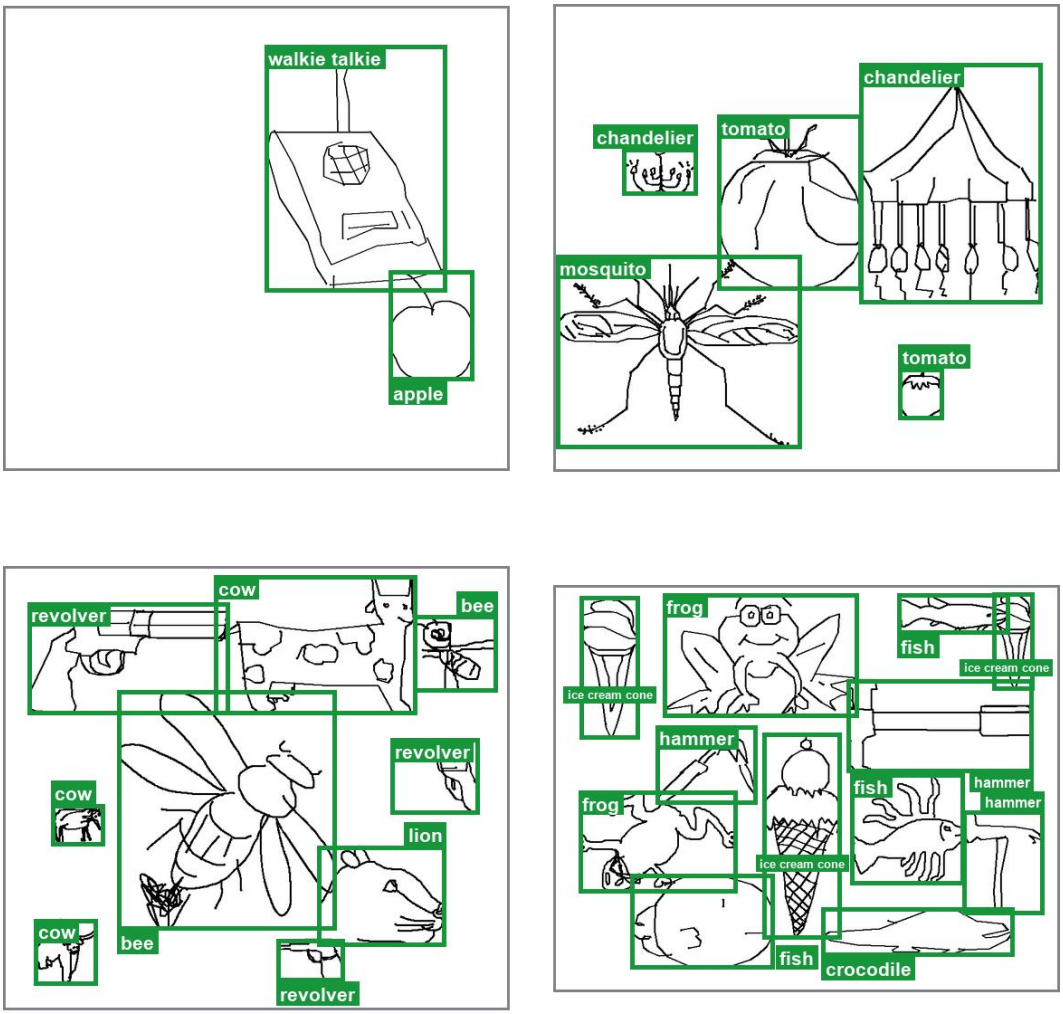

Figure 7: **Composited free-hand sketch scenes.** Four scenes with their ground-truth boxes, ordered by instance count. Canvas size and aspect ratio are sampled per scene.

Sketchy stitched canvases for Stage B. Sketchy provides photo-sketch pairs but no bounding boxes, so each photo receives a tight box from GroundingDINO (Liu et al., 2024) prompted with its category name. A square canvas with side sampled between 768 and 1280 pixels tiles 1 to 7 such photos drawn from up to 4 categories (Fig. 8). Instance-level samples place several tiles of the queried category, so the query competes with a same-category distractor. Tiles are cropped off-centre within the margin around the tight box, resized across a stratified size range, colour jittered,

alpha-feathered at their borders, and placed so that no tile overlaps a previously placed object region. Every ground-truth box therefore stays fully visible. Backgrounds are Objects365 photographs screened to contain none of the pasted categories. Query sketches are inverted to white-on-black so their polarity matches the HED sketches of Stage A.



Figure 8: **Stitched Sketchy canvases.** Each group holds the query sketch and the canvas it is asked about. Green marks the queried instance, orange a same-category distractor, and grey every other object.

Objects365 HED-sketch queries for Stage B. For the Objects365 augmentation, HED is applied to the whole photograph, and the query is cropped from the resulting edge sketch at one annotated object's box, with a random padding of 2.5 to 7.5% of the box size per side. Boxes with area below 96×96 pixels are skipped. The model is then asked to localise that instance in the untouched source photograph (Fig. 9).

SketchyCOCO overlap filter. SketchyCOCO's foreground sketches were themselves retrieved from QuickDraw, TU-Berlin and Sketchy, the three pools the constructions above draw from, so an evaluation query could otherwise appear among the training drawings. Two levels of filtering separate them. At the category level, every Sketchy pair whose category is one of SketchyCOCO's 14 foreground categories is dropped from training: airplane, bicycle, car, cat, cow, dog, elephant, fire hydrant, giraffe, horse, motorcycle, sheep, traffic light and zebra. At the instance level, every SketchyCOCO evaluation query is matched against all three pools and every drawing it matches is denied to both constructions. A drawing's signature is its ink cropped to the tight box, stretched to 32×32, Gaussian blurred and L2 normalised, and a candidate is denied once its cosine similarity to a query reaches 0.92, with the horizontal flip of each query matched as well. The stretch removes the aspect distortion SketchyCOCO applies when fitting a sketch into an instance box, and the blur absorbs the stroke-width and antialiasing differences between its renders and ours.

C FS-COCO Annotation

Procedure. Two people annotated the official FS-COCO test split and a third checked every annotation, using a browser tool built for the purpose. It shows the photograph, with COCO's boxes

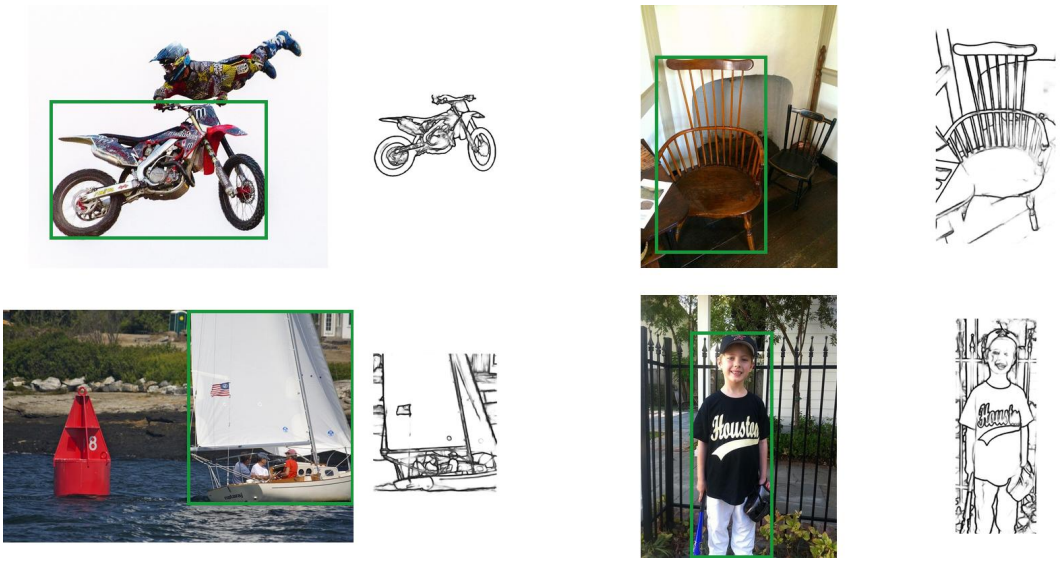

Figure 9: **Objects365 HED-sketch queries.** Each group holds a photograph with the queried box and the HED-sketch crop taken from that box, which serves as the query. Sketches are drawn black on white here, and training uses the inverted polarity.

drawn on it, beside the rendered sketch, and records every box and link as it is drawn, so each link is made with both sides in view.

Boxes on the sketch. Labels come from the COCO 80 categories extended with five that appear in these sketches but COCO does not name: tree, building, cloud, road sign, and tower. Every sketched object of one of these categories receives a box, tightened to the outer edge of its strokes. A box is drawn only when the strokes inside it are recognisable as that object on their own, without looking at the photograph. The ground truth is therefore self-contained.

Sketch-photo links. The photograph's boxes are COCO's own annotations and stay untouched, so linking runs in one direction only. A sketched object is linked to a photographed instance when the two depict the same object, and the link is kept only when the sketch crop alone identifies that instance uniquely. An instance annotated in the photograph that also appears in the sketch receives a box on the sketch side and is then linked, while a sketched object with no counterpart among the photograph's annotations stays unlinked and serves the detection and pointing ground truth alone. A sketched object whose category COCO does not annotate therefore carries no query. This yields the 2754 query-target pairs of Sec. 4.1.

Category distribution. The annotation holds 11674 boxes over 68 categories on 2907 sketches, 4.0 boxes per sketch, and the queries span 59 of those categories. Queries per category have a median of 15 against a mean of 47. The distribution is long-tailed (Fig. 10), as people drawing a scene from memory return to a few familiar objects far more often than to the rest.

D Affordance Evaluation Protocol

AGD20K supplies a heatmap for each image and affordance verb, and its official metrics compare predicted and ground-truth heatmap distributions. Every model in Table 5 emits points rather than a heatmap, so we score containment instead. The heatmap is binarised at half its peak value. Each connected region of the result becomes one ground-truth target, since one verb can apply to several instances in an image. Predicted points are matched to these targets one-to-one and greedily, as boxes are matched in Sec. 4.1, and a match holds when the point falls inside its target. The label a model attaches to a point is ignored, so extra points count against precision. Recall and precision are computed over all matches, and F1 is their harmonic mean.

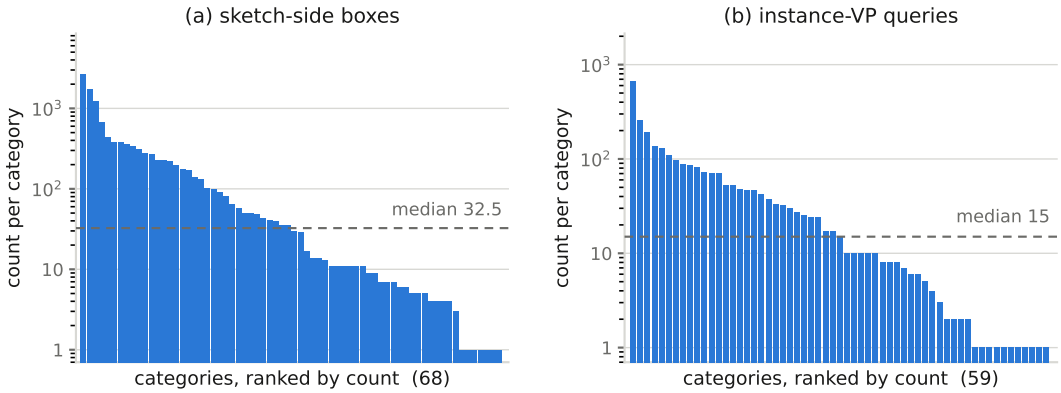

Figure 10: **Per-category counts of the FS-COCO annotation.** Categories are ranked by count and the vertical axis is logarithmic. The dashed line marks the median.

Table 6: **Stratified instance VP, all baselines.** Each cell is SketchyCOCO / FS-COCO, same protocol as Table 2.

Method	Subset	F1@0.3	F1@0.5	F1@0.7
Qwen2.5-VL-3B	Easy	75.7/57.4	70.3/52.6	64.9/48.8
	Hard	42.3/43.5	36.3/39.1	32.1/37.2
Qwen3-VL-8B	Easy	91.0/86.9	**90.2**/84.3	**86.7**/79.5
	Hard	50.5/66.2	47.8/63.1	43.4/56.8
MiMo-VL-7B	Easy	83.9/51.9	82.9/48.9	79.0/43.0
	Hard	47.3/30.5	43.8/28.9	39.8/26.1
DeepSeek-VL2-S	Easy	65.2/45.1	62.8/37.7	58.3/27.0
	Hard	42.0/30.3	39.1/25.0	35.9/19.3
Sketch-Omni	Easy	**91.7/94.8**	89.5/**92.7**	86.1/**88.3**
	Hard	**55.3/83.3**	**51.4/80.3**	**48.2/77.0**

Table 7: **Stage-A alignment on the sketch-native tasks.** Detection and referring report F1@0.5, pointing F1@Point, affordance F1, all under the protocol of the main tables. Referring uses the test splits.

Method	Detection		Pointing		Referring		Affordance	
	FS-COCO	COCO	FS-COCO	COCO	obj	spat	Seen	Unseen
Sketch-Omni (full)	86.7	48.1	89.9	44.5	74.6	76.7	66.2	74.3
w/o align	86.3	45.2	89.2	42.4	70.3	73.1	63.9	70.8

E Stratified Instance VP Across All Baselines

Table 6 extends the stratification of Sec. 4.3 to every MLLM baseline. The split yields 689 Easy and 651 Hard queries on SketchyCOCO, and 2452 Easy and 302 Hard queries on FS-COCO. Base Rex-Omni performs near floor on this task, as Table 1 shows, and is omitted. Every model loses F1 from Easy to Hard. Sketch-Omni leads every MLLM baseline on both Hard subsets and on FS-COCO-Easy, and is on par with Qwen3-VL-8B on SketchyCOCO-Easy.

F Stage-A Alignment on the Sketch-Native Tasks

Table 7 evaluates the *w/o align* configuration of Table 1 on the five sketch-native tasks, isolating the alignment's effect on the tasks it operates on during Stage A. The largest gain falls on object referring, 4.3 points. This is consistent with the role of $\mathcal{L}_{\text{cross}}$, which aligns how a phrase selects its region.

G Training Details

Supervised stages. Both stages fit the LoRA update $\Delta\theta$ with a single AdamW optimiser at an initial learning rate of 5×10^{-5} with cosine decay, weight decay 0.01, and gradient-norm clipping at

1.0, at an effective batch size of 64. The adapters use scaling factor 64 and no dropout. Training runs in bfloat16 with gradient checkpointing under ZeRO stage 1, for about 1 day on 32 H100 GPUs.

Task mixture. Stage A and Stage B alternate in blocks of 50 optimiser steps, two Stage-A blocks to every three Stage-B blocks. Within a block the tasks are sampled at the rates below, which give the share of all optimiser steps in the last column. Object and spatial referring share the referring stream and together receive probability 0.25.

Stage	Task	Within stage	Overall
A	detection	0.25	0.10
A	pointing	0.25	0.10
A	referring	0.25	0.10
A	affordance	0.25	0.10
B	instance VP, stitched canvases	1/3	0.20
B	instance VP, crop queries	1/3	0.20
B	composited-scene detection	1/6	0.10
B	composited-scene pointing	1/6	0.10

GRPO refinement. The final GRPO pass follows Rex-Omni's recipe and runs for about 8 hours on 32 H100 GPUs on the same task mixture as the supervised stages, with 8 rollouts per prompt at temperature 1.0 and top-p 0.99, a global batch size of 32, learning rate 5×10^{-6}, PPO clip 0.2, and KL coefficient 0.01. Box tasks are rewarded by the IoU between the predicted and ground-truth boxes, and pointing tasks by whether the predicted point lands inside the ground-truth box or mask.