What Sketches Tell Us about LVLMs: Conventions, Grounding, and Localisation

Chaitat Utintu[1,2], **Ahmed Bourouis**[1], **Arkaprabha Basu**[1], **Yi-Zhe Song**[1,2]
[1] SketchX, Surrey Institute for People-Centred Artificial Intelligence (PAI), University of Surrey
[2] UKRI AI: CDT in AI for Digital Media Inclusion
{chaitat.u, a.bourouis, a.basu, y.song}@surrey.ac.uk

Abstract

LVLM benchmarks tell us whether a model answers correctly, but not which visual conventions it has retained, nor where those conventions live. Sketches expose this hidden structure. A sketch removes texture, lighting, and photographic context while preserving the strokes humans judged sufficient for recognition. If an LVLM can operate on that abstraction, the surviving signal is structural rather than merely photometric. We introduce a training-free sketch probe with three readouts: temporal part-construction order, spatial attention-based grounding, and the mechanistic layer/head locus of grounding. We evaluate nine LVLMs on part order and five on the attention readouts, over two part-annotated sketch datasets. Three findings emerge. First, LVLM part order tracks human convention: models reproduce canonical orderings where human drawings converge, and become correspondingly variable where human orderings vary. Controls that blank, withhold, and mismatch the sketch locate this convention in the language model. Second, part-level grounding survives the abstraction: attention heads selected from the frozen model with an existing recipe recover part and object regions on object and scene sketches, with no grounding supervision. Third, the grounding signal concentrates in a narrow mid-network band whose location follows the base language model rather than parameter count in the two families we compare, and remains stable across object and scene sketches. The result is a practical diagnostic for sketch-driven LVLMs: no fine-tuning, no auxiliary segmenter, and a layer band that doubles as a head-selection prior, matching or improving grounding mAcc@0.1 in every model and dataset we tested.

1 Introduction

LVLMs can recognise, describe, and even structure familiar objects with striking competence. A coffee cup gets a handle, an opening, and a body in roughly the right arrangement. Standard multimodal benchmarks [37, 26] score the final answer, but leave a central question unresolved: which visual conventions has the model retained, and where in the network are they carried?

This question matters because LVLMs [35, 7, 43, 3, 20] are trained on human-produced images and language. Their object knowledge is therefore entangled with the conventions humans use when drawing, depicting, and naming visual structure. A model may answer a part-level question about a photograph correctly by leaning on texture, colour, background, or dataset co-occurrence, without using the part-level schema of the object at all. By a *visual convention* we mean a regularity in how humans depict a concept that is not entailed by its physical structure, such as which parts are drawn and in what order: that a cup has a handle is a fact about cups, whereas drawing the body before the handle is a convention.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

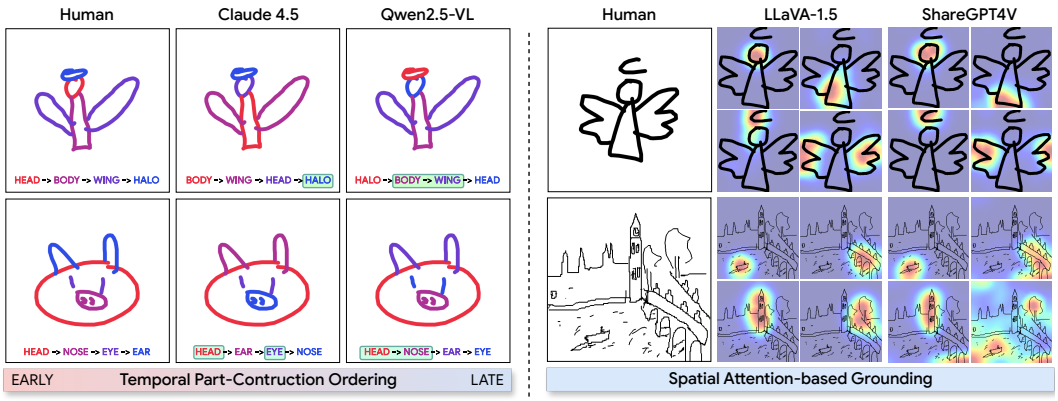

Figure 1: **Sketches as a probe for LVLMs.** Sketches keep only the strokes humans considered essential, exposing what an LVLM treats as essential. We probe this from three complementary perspectives. **Left**: *temporal* part-construction order on two object sketches (early-to-late gradient, predicted ordering beneath) against human consensus. **Right**: *spatial* attention-based grounding on one object and one scene sketch, with per-part / per-object overlays beside the human sketch.

Reading these out requires a probe that strips surface signal whilst preserving structural commitments. Natural images are the wrong probe: photographs carry texture, lighting, and contextual scaffolding that the model can lean on, masking whatever schematic structure it has acquired. Colour and texture are correlates of object identity that do not require decomposing an object into parts, whereas spatial arrangement is constitutive of part-level structure, and it is the channel for which SPG [34] records human ground truth stroke by stroke. A sketch keeps only the strokes a person judged essential and discards everything else, so what survives is the part-level structural skeleton of the depicted concept. Probing an LVLM with sketches asks the model to operate on this skeleton alone, and what it surfaces is a direct readout of the visual conventions it has retained (Fig. 1).

What this paper claims. Such a probe should also read internal state rather than output accuracy. Ours is a training-free diagnostic, deliberately assembled from existing components: shuffled-list ordering for part-construction order, the Localization Heads recipe of Kang et al. [30] for grounding-head selection, and classical sequence metrics [32, 29, 31] for trajectory comparison. The analytical weight therefore falls on the frozen LVLMs, not on new apparatus.

We instantiate three readouts on two part-annotated sketch datasets [34, 19]: temporal part-construction order across nine LVLMs, three of them closed-source, and spatial attention-based grounding and its mechanistic layer/head locus across five open models on two visual stacks. The readouts expose two facts about *what* survives abstraction and one about *where* it is stored:

(i) *What conventions persist.* LVLMs reproduce canonical part-orderings on categories where human drawings converge, and become correspondingly unstructured where they do not, deferring to human convention rather than imposing their own schema. Controls that blank, withhold, and mismatch the sketch place this convention in the language model itself.

(ii) *What is grounded without training.* Heads selected with the recipe of Kang et al. [30] recover part and object regions on both object and scene sketches from frozen cross-attention alone. Against the trained grounders CLIPSeg [40] and GroupViT [59], they overlap the target region more often at coarse and moderate thresholds, so part-level grounding is present in the representation rather than added by fine-tuning. At scale the two readouts couple class by class (LLaVA-1.5-13B Pearson $r = +0.40$).

(iii) *Where the apparatus lives.* Grounding heads concentrate in a narrow mid-network band whose location tracks the base language model across the two families we compare (Vicuna at fractional depth 0.35 to 0.45, DeepSeek at 0.50 to 0.65) and is conserved across object/scene domains ($\Delta = +0.004$). The same heads recur across domains, and ablating the three most frequently selected ones lowers the margin towards the correct answer on a downstream localisation question by 0.118, against 0.010 for heads drawn from the same layers.

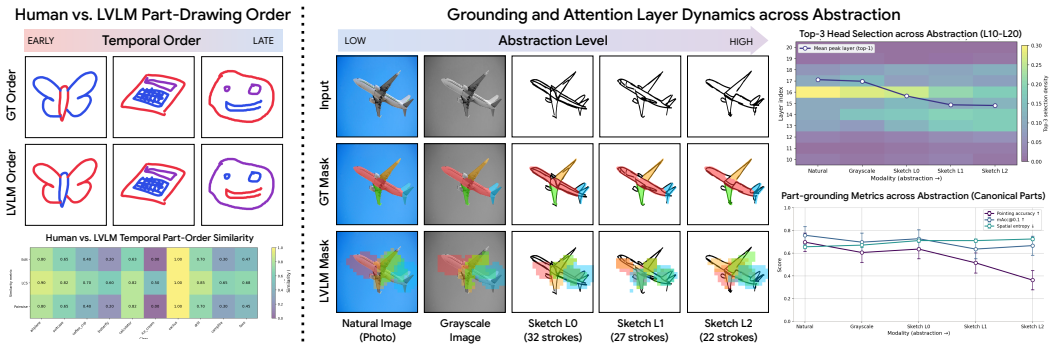

Figure 2: **Pilot: part-construction ordering and grounding across abstraction. Left**: human (GT) and LVLM part-drawing trajectories on three SPG [34] categories (early→late coloured), with mean per-sketch part-order similarity ($\pm$ SD, $N = 100$) across ten categories. **Right**: LVLM part-localisation on Pascal-Part [15] across ControlSketch [4] abstraction levels, top-3 grounding-head selection (peak shifts to earlier layers), and degrading pointing accuracy and mAcc@0.1 with stable spatial entropy. See Sec. 3.1.

What these insights enable. Insight *(ii)* is the most immediately actionable: sketch-based retrieval [8], segmentation, editing, and AR/UI pipelines that currently bolt on a dedicated segmenter could read region grounding directly from an LVLM they already run. Insight *(iii)* has implications for backbone selection in grounding-heavy work: band position follows LM lineage rather than parameter count in the families we compare, so the smaller variant in the right family often suffices, and a band that is causally load-bearing is a natural intervention point for attention-steering methods [61] that would otherwise have to search the network blind.

Contributions. We make three contributions. (i) We introduce a training-free sketch probe with temporal, spatial, and mechanistic readouts of LVLM internal organisation, computed from frozen models. (ii) We show that LVLMs track human sketch conventions rather than imposing a fixed model schema, with controls that locate this convention in the language model, and that selected attention heads recover part and object regions on object and scene sketches without grounding supervision. (iii) We localise the grounding signal to a narrow family-dependent mid-network band, confirm by ablation that the selected heads carry it, and show that restricting head search to this band acts as a usable prior, matching or improving grounding mAcc@0.1 across every model and dataset tested.

2 Related Work

Interpretability in Large Vision-Language Models. Visual model interpretability has evolved from CNN saliency [62, 64, 53] through transformer attribution [1, 13, 2, 52] to self-supervised ViTs [22] and contrastive VL models [47]. With instruction-tuned LVLMs [36, 35, 5, 33], grounding can be generated via location tokens [44], unified with referring [60], or elicited from LLM-style decoders [48], while specific attention heads exhibit informative direction [12, 30] and visual prompts can steer prediction [61, 54]. Object information stays spatially localised across visual tokens, progressively aligning to interpretable vocabulary [42]. Yet these studies probe natural images, where CLIP-based encoders harbour systematic blind spots [55]. The word grounding also covers three distinct mechanisms in this literature: generative grounding, where the model emits location tokens or referring expressions [44, 60], gradient and attribution saliency [13, 53, 62, 64], and attention-based localisation, where the region is read directly from attention weights [12, 30]. Ours is strictly the third, with no coordinate decoding, no gradients, and no auxiliary segmenter. We use sketches as a controlled abstract input, asking whether attention-based localisation survives when texture, lighting, and context are absent, and reporting both its accuracy and the layer band that carries it.

Sketch understanding and interpretability. Sketches are a sparse, abstract medium long used for recognition [23, 28, 49], retrieval [8, 51], grouping [45, 34], and generation [58, 27]. Optimisation-based methods produce object [57] and scene sketches [56], and benchmarks [11, 19, 17] evaluate LVLMs on line drawings, but treat sketches as a domain shift rather than a probe of internal representation. Sketch-specific explainability [46], saliency [9], and detection [18] exploit this further, while cognitive studies use sketches as a window onto visual concept organisation [25, 38, 41],

recovering trajectories that reveal stable construction orders [24]. How LVLMs internalise comparable structure remains untested. We treat the sketch itself as the experimental probe, turning sketches from a domain to evaluate *on* into a probe to evaluate *with*.

3 Method

In this section, we first present a **pilot study** in Sec. 3.1 examining part-construction ordering and grounding across visual abstraction, motivating attention as the central probe signal. We then formalise our **three-perspective probing framework** in Sec. 3.2, treating the temporal, spatial, and mechanistic axes as complementary readouts of internal representation. Finally, Sec. 3.3 instantiates each perspective as a training-free probe over object and scene sketches.

3.1 Pilot study: part-construction ordering and grounding under abstraction

Before instantiating our framework at scale, we conduct a pilot study to motivate its central claim: that the part-level structure an LVLM has internalised is revealed most sharply under abstraction, where stripping texture and context exposes its structural commitments. Fig. 2 summarises the findings. *Left (temporal):* we contrast human part-drawing trajectories from SPG [34] with LVLM-induced orderings on the same sketches. Both follow coarse-to-fine sequences, and the heatmap reports per-sketch Edit, LCS, and Pairwise similarity (mean $\pm$ SD over 100 sketches per category) across the ten categories whose human drawings most consistently follow a shared canonical order. Agreement is strong on canonical categories such as *calculator* and *coffee_cup* but degrades on parts where humans themselves disagree, and within-class variance surfaces categories where the LVLM produces near-random orderings rather than a single canonical guess. *Right (spatial & mechanistic):* we probe attention-based part grounding on Pascal-Part [15] across natural photo, grayscale, and three ControlSketch [4] abstraction levels. This is a graded ablation over the surface channels: grayscale removes colour while keeping texture and context, and the sketch levels then remove texture and context progressively. LVLM masks track ground truth at low abstraction but degrade as input becomes sparser. The peak grounding layer migrates earlier with increasing abstraction, and pointing accuracy and mAcc@0.1 decline substantially whilst spatial entropy rises only modestly. The model becomes incorrect about *where* parts lie whilst its attention shape changes only slightly. Abstraction therefore selectively erodes spatial accuracy whilst preserving temporal structure and broad attention shape, motivating the probes formalised in Sec. 3.3.

3.2 Problem formulation and method overview

Problem. We treat the sketch input as a controlled stripping of texture, lighting, and surface context that exposes what an LVLM has internalised about objects and their parts. Consider a pre-trained LVLM with L transformer layers and H heads per layer. The input is a pair (I, p), where I is a sketch (object or scene) and p identifies a target part or whole object. A prompt template $T(p)$ (Sec. 3.3) converts p into a textual probe query, after which we read out the cross-attention $\alpha_{\ell,h}(p) \in \mathbb{R}^{P^2}$ at each layer ℓ and head h, where P^2 is the number of image patches. The model is used purely via forward passes, with no fine-tuning, auxiliary networks, or external segmentation. Our goal is to characterise the part-level structure that persists in the LVLM's internal state under sketch input. The pilot study (Sec. 3.1) reveals three signals to formalise: *temporal* part-construction order, *spatial* attention-based grounding, and *mechanistic* layer/head locus.

Three readouts. Fig. 3 illustrates the framework. Rather than a single aggregate score, it returns three readouts of internal state from forward passes of one frozen LVLM on (I, p). The *temporal* readout records the order $\pi(p)$ in which the model returns a shuffled part list and compares it with human drawing trajectories by edit distance, longest common subsequence, and pairwise-order agreement. The *spatial* readout selects a small set $\mathcal{G}$ of grounding heads by image-attention mass and spatial concentration, and turns their combined attention into a part mask. The *mechanistic* readout characterises $\mathcal{G}$ itself, recording the layer and head of each selected head and the spatial entropy of the combined map, and asks whether the band that carries grounding is conserved across models and across the object and scene divide. Sec. 3.3 defines each in turn.

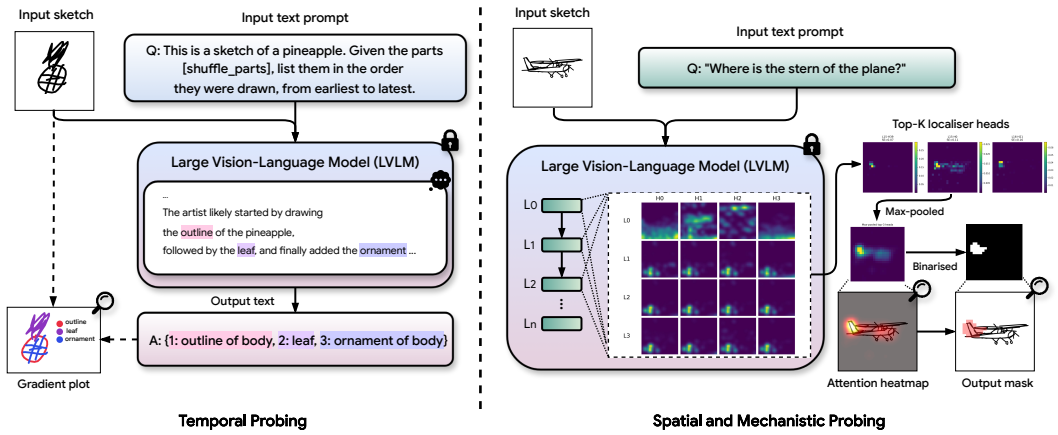

Figure 3: **Three-perspective probing framework.** The same frozen LVLM is queried twice on a single sketch under two protocols. **Left (temporal probing):** the model receives the sketch with a shuffled set of parts and is asked to return them in drawing order. **Right (spatial and mechanistic probing):** the model is asked to locate a named part, and attention is filtered to the top-k grounding heads and max-pooled into a grounding mask.

3.3 Sketch-based probes of part-construction ordering and grounding

Temporal probing. We elicit a part-construction trajectory $\pi(p) = (p_1, \ldots, p_K)$ in a single forward pass: the model is shown the sketch, the object name, and the part list shuffled into set notation $\{p_1, \ldots, p_K\}$, and is asked to return the parts in the order they were drawn. The protocol is purely textual on the parts side: no sketch is re-rendered, so the resulting sequence reflects only the model's internal ordering rather than feedback from a partial drawing. Sec. 4.2 tests whether the sketch image itself drives this ordering. We compare $\pi(p)$ against human trajectories π^{human} from SPG [34], which records per-sketch part strokes and yields a stable modal trajectory per category, using edit distance, longest common subsequence, and pairwise-order agreement (Kendall's τ [31]), which jointly capture absolute position, the longest shared ordered subsequence, and the relative precedence of part pairs. The parts vocabulary and prompt template are held fixed across all sketches, so any variation in $\pi(p)$ is attributable to the model. The temporal probe is restricted to objects, since SPG supplies human orderings only at the part level within a single object.

Spatial and mechanistic probing. We query each input with a template $T(p)$, i.e., *"Where is the {p} of the {c}?"* for object parts on SPG [34] and *"Where is the {c}?"* for whole objects on FS-COCO [19], and extract cross-attention from the *last input token*, which under causal masking aggregates information across the whole prompt. Let $\alpha_{\ell,h}(p) \in \mathbb{R}^{P^2}$ denote this attention vector at layer ℓ, head h, restricted to the P^2 image patches. Following Kang et al. [30], who validated the recipe on ten LVLMs including all five used here, we score each head by image-attention mass $\rho_{\ell,h}(p) = \sum_{i \in \mathcal{V}} \alpha_{\ell,h,i}(p)$ over the visual tokens $\mathcal{V}$ and spatial concentration, measured by the spatial entropy $\eta_{\ell,h}(p)$ of the attention map (lower is more concentrated):

$$\eta_{\ell,h}(p) = -\sum_{i=1}^{P^2} \tilde{\alpha}_{\ell,h,i}(p) \log \tilde{\alpha}_{\ell,h,i}(p), \quad \tilde{\alpha}_{\ell,h}(p) = \alpha_{\ell,h}(p) \,/\, \textstyle\sum_i \alpha_{\ell,h,i}(p). \tag{1}$$

Heads are scored jointly by these two criteria, reduced via the chord-elbow rule of [30], and the top $k = 3$ form the grounding-head set $\mathcal{G} \subset \{1, \ldots, L\} \times \{1, \ldots, H\}$. Selected heads are smoothed (mean-subtract, ReLU, Gaussian), combined by element-wise maximum, and thresholded at the 80th percentile to yield a mask $M(p)$, scored against ground-truth part masks (SPG) and object masks (FS-COCO) via pointing accuracy, mAcc@0.1, and spatial entropy (SE). The mechanistic readout reuses the same machinery: we record the (layer, head) coordinates of $\mathcal{G}$ and the per-sample SE, and ask how their distributions vary across models and across the object/scene divide. All three probes require no fine-tuning, no auxiliary heads, and no rendering between iterations.

4 Experiments

4.1 Model sets, protocols, and why they differ across probes

The three probes draw on overlapping but distinct model sets, dictated by the methodology each requires. The temporal probe needs only generated text, so it also covers closed-source models, whereas the attention readouts need access to the attention weights. For temporal probing (Tab. 1) we use six open-source models spanning four families (LLaVA-1.5-7B/13B [35], Qwen2.5-VL-3B/7B [7], Molmo-7B-D [21], InternVL2-8B [16]) and three closed-source models from the smallest tier of each frontier family (GPT-5 (nano) [43], Claude Haiku 4.5 [3], Gemini 2.5 Flash-Lite [20]). All models use deterministic decoding (temperature 0), and closed-source predictions are averaged over three runs per item for stability. Trajectories are elicited on SPG [34] (25 categories) under a class-consensus protocol: for each category we identify the modal human trajectory and sample 100 strict-mode-matching sketches, presenting them to each LVLM with the part list shuffled and displayed in set notation $\{p_1, \ldots, p_K\}$ to neutralise input ordering bias. Per-class scores aggregate over the 100 sketches, and we report mean $\pm$ SD across the 25 categories to expose inter-class variance, which is itself a finding. For attention-based grounding (Tab. 2) and the mechanistic profile (Fig. 6) we use five open models. LLaVA-1.5-7B/13B and ShareGPT4V-7B [14] use a CLIP ViT-L/14 vision encoder and a Vicuna language model, whereas DeepSeek-VL-1.3B/7B [39] pair a SigLIP-based vision encoder and their own image processor with DeepSeek-LLM. The set therefore spans two distinct visual stacks and two base language model families, and the Localization Heads recipe [30] transfers to both unchanged. Cross-attention $\alpha_{\ell,h}(p)$ from the last input token to the P^2 image patches is extracted across all L layers and H heads, smoothed per head (mean-subtract, ReLU, Gaussian $\sigma = 1$, normalise), and aggregated over the top-$k = 3$ heads by image-attention mass and spatial concentration. Grounding is evaluated on SPG (object) and FS-COCO [19] (scene), with FS-COCO scene-level segmentation masks taken from FS-COCO-Seg [11], against CLIPSeg [40] and GroupViT [59], with dataset-level means reported from a single forward pass with no fine-tuning.

4.2 Main results

Part-construction order. Before contrasting models, we examine the human side of SPG. Fig. 4 ranks all 25 categories by strict-mode consensus, and reveals a strongly bimodal distribution: canonical objects such as *airplane*, *suitcase*, and *coffee_cup* elicit near-unanimous part orders ($\geq$ 90%), whereas creature-like or symmetry-rich categories such as *angel*, *pig*, and *house* fall below 40%, indicating that humans themselves do not converge on a single drawing schema. This dataset-side heterogeneity sets the bar against which any LVLM prediction must be read. Tab. 1 reports the resulting agreement under all three metrics. The closed-source row leads as expected, exceeding every open-source model. Within the open-source families, however, the gap is markedly narrower than on standard multimodal benchmarks: Qwen2.5-VL-7B reaches 0.84 LCS, within 0.05 of the closed-source leader, suggesting that part-construction order is a relatively shallow inference once a model has the object's name and its part list. Newer open models narrow the gap further (App. H). The per-class view in Fig. 5 (left) reveals where the alignment holds: categories with strong human consensus (*calculator*, *campfire*, *ice_cream*) cluster near the diagonal with high LVLM agreement, whereas categories where humans themselves disagree (*angel*, *pig*, *flower*) drop into the lower-left, with LVLM agreement degrading commensurately. LVLMs internalise the canonical part schemas humans use when those schemas are stable. Where human consensus dissolves, model behaviour becomes correspondingly unstructured rather than collapsing onto a single canonical guess. Across the 25 categories, per-category agreement correlates with human consensus (mean Pearson $r = +0.58$ over all nine models, significant in eight, App. B). Memorisation of SPG would not produce this pattern, since a stored ordering would be reproduced equally well where humans disagree.

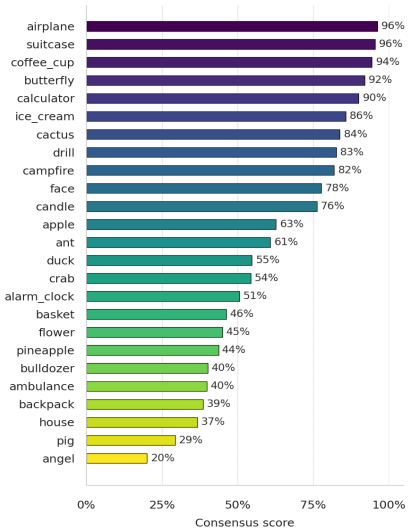

Figure 4: Human consensus statistics across all SPG [34] sketch categories.

Table 1: **Part-construction order on object sketches.** LVLM trajectories $\pi(p)$ compared against human trajectories π^{human} on SPG [34] under the class-consensus protocol, evaluated with edit distance [32], longest common subsequence [29], and pairwise-order agreement (Kendall's τ) [31]. Best per column is highlighted within closed-source (purple) and open-source (blue).

Model	SPG [34]		
	Edit↑	LCS↑	Pair↑
Closed-source LVLMs			
GPT-5 (nano) [43]	0.72 (±0.26)	0.85 (±0.14)	0.81 (±0.19)
Claude Haiku 4.5 [3]	**0.80 (±0.20)**	**0.89 (±0.11)**	**0.88 (±0.13)**
Gemini 2.5 Flash-Lite [20]	0.73 (±0.25)	0.85 (±0.14)	0.83 (±0.17)
Open-source LVLMs			
InternVL2-8B [16]	0.66 (±0.28)	0.80 (±0.18)	0.72 (±0.28)
Molmo-7B-D [21]	0.63 (±0.24)	0.78 (±0.16)	0.71 (±0.22)
Qwen2.5-VL-3B [7]	0.57 (±0.26)	0.77 (±0.14)	0.64 (±0.24)
Qwen2.5-VL-7B [7]	**0.69 (±0.22)**	**0.84 (±0.12)**	**0.80 (±0.15)**
LLaVA-1.5-7B [35]	0.38 (±0.24)	0.66 (±0.12)	0.46 (±0.26)
LLaVA-1.5-13B [35]	0.46 (±0.16)	0.69 (±0.09)	0.51 (±0.18)

Table 2: **Attention-based grounding on object and scene sketches.** Cross-attention grounding accuracy of LVLM grounding heads on object (SPG) and scene (FS-COCO) sketches, under pointing accuracy [63], mAcc@0.1 [10], and spatial entropy [50]. Best per column is highlighted within trained grounders (purple , shown for reference) and LVLMs (blue).

Model	Object, SPG [34]			Scene, FS-COCO [19]		
	Point↑	mAcc@0.1↑	SE↓	Point↑	mAcc@0.1↑	SE↓
Trained grounders						
CLIPSeg [40]	**0.56**	0.27	**0.96**	**0.67**	0.40	**0.96**
GroupViT [59]	0.34	**0.59**	0.98	0.39	**0.60**	0.97
Training-free LVLM probes						
ShareGPT4V-7B [14]	0.42	0.72	0.87	0.45	0.72	0.86
DeepSeek-VL-1.3B [39]	0.31	0.56	0.92	0.34	0.61	0.91
DeepSeek-VL-7B [39]	0.23	0.48	0.85	0.33	0.65	0.84
LLaVA-1.5-7B [35]	0.43	0.65	0.88	0.53	0.69	0.87
LLaVA-1.5-13B [35]	**0.46**	**0.73**	**0.84**	**0.61**	**0.78**	**0.83**

Does the sketch drive the ordering? As the prompt supplies the object name and its parts as text, we test what the image adds. With prompt, parts, shuffle, and decoding fixed, we replace the sketch with a blank canvas, remove it (Tab. 3), or substitute a sketch of another category. Of eighteen paired comparisons over the 25 categories (two models, three conditions, three metrics), none favours the real sketch, and the only significant one favours the no-image condition (Pair, $p = 0.04$). A test on sketches whose human order departs from the canonical one ($p = 0.26$ under clustering by category) and a forced choice with chance at 50% by construction agree (App. C). The ordering convention is therefore carried by the language model. SPG remains the instrument that measures it, and the spatial and mechanistic readouts, which read attention over image patches, are unaffected.

Attention-based grounding. We turn from temporal ordering to where the LVLM places its attention. Tab. 2 reports cross-attention grounding accuracy on object (SPG) and scene (FS-COCO) sketches under pointing accuracy, mAcc@0.1, and spatial entropy. The strongest LVLM probe, LLaVA-1.5-13B, reaches 0.46 Point and 0.73 mAcc@0.1 on SPG, and 0.61 Point and 0.78 mAcc@0.1 on FS-COCO. The comparison against the trained grounders depends on what is measured. LLaVA-1.5-13B exceeds GroupViT [59] on Point but trails CLIPSeg [40] (0.46 vs 0.56 on SPG, 0.61 vs 0.67 on FS-COCO), whereas on mAcc@0.1, which asks only whether the predicted mask overlaps the target region, it leads both, with CLIPSeg at 0.40 on FS-COCO against 0.78. At stricter thresholds

Table 3: **Image controls on the temporal probe.** 50 sketches per category, prompt and shuffles fixed. Real-sketch rows are within 0.01 of Tab. 1.

	Qwen2.5-VL-7B			LLaVA-1.5-7B		
Input	**Edit↑**	**LCS↑**	**Pair↑**	**Edit↑**	**LCS↑**	**Pair↑**
Real sketch	0.69	0.84	0.79	0.37	0.66	0.46
Blank canvas	0.69	0.84	0.79	0.39	0.68	0.50
No image	0.72	0.86	0.83	0.44	0.69	0.51

Table 4: **Grounding at stricter overlap.** All methods re-scored at IoU 0.3 and 0.5. mAcc@0.1 is in Tab. 2.

	SPG		FS-COCO	
Model	**@0.3↑**	**@0.5↑**	**@0.3↑**	**@0.5↑**
CLIPSeg [40]	0.17	0.12	0.31	0.20
GroupViT [59]	0.28	0.11	0.21	0.08
LLaVA-1.5-13B	0.29	0.06	0.35	0.11

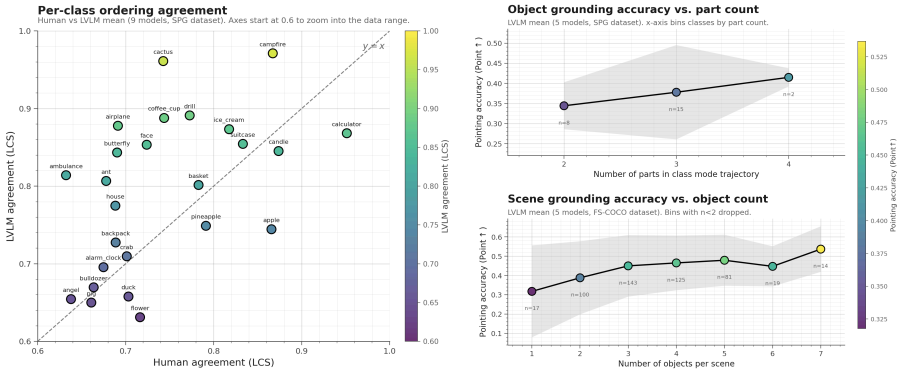

Figure 5: **Summary of ordering and grounding behaviour. Left**: per-class human-LVLM ordering agreement under LCS [29], with viridis colour encoding LVLM agreement. **Right**: LVLM grounding accuracy as a function of input complexity, on object (SPG, parts) and scene (FS-COCO, objects).

(Tab. 4) the lead holds at mAcc@0.3, significant in three of four comparisons, and is gone at mAcc@0.5, where the trained grounders are nominally ahead in three of four comparisons, none significant. Frozen cross-attention therefore matches or exceeds them at coarse and moderate overlap, while at strict overlap its fixed-area mask covers only about half of the part (App. D). Within the LLaVA family, scale yields a consistent gain (0.43→0.46 Point on SPG, 0.53→0.61 on FS-COCO), yet within DeepSeek-VL the larger 7B variant underperforms the 1.3B on objects (0.23 vs 0.31 Point), indicating that grounding quality is not a monotonic function of parameter count but depends on training-data composition and instruction-tuning recipe. Fig. 5 (right) makes the input-side dependency explicit: scene-grounding Point on FS-COCO rises from 0.32 at one object to 0.55 at seven objects, and a similar though smaller trend (0.34→0.42 from two to four parts) holds within SPG. Richer compositional context supplies the cross-attention with additional disambiguating evidence, sharpening the localisation peak. The SPG trend is the cleaner signal, since each query targets a single part whose mask area does not grow with part count, ruling out a coverage artefact. This pattern mirrors the pilot (Fig. 2, right), where increasing abstraction along the natural-photo → grayscale → sketch axis monotonically degraded grounding: at one end, stripping context erodes accuracy, and at the other, denser composition restores it. LVLM grounding therefore behaves as a graded function of input richness, not a brittle binary of in- versus out-of-domain.

Mechanistic profile. Having established that frozen cross-attention recovers part regions, we now ask *where* in the network this signal lives. Fig. 6 (top) shows that the top-3 grounding heads concentrate in a narrow mid-network band in every model, but the band's location is not uniform: the three Vicuna-based models (LLaVA-1.5-7B/13B, ShareGPT4V-7B) place selection mass around fractional depth 0.35 to 0.45, whereas the two DeepSeek-based models sit visibly deeper, with peaks in 0.50 to 0.65. Crucially, the band is conserved when the input shifts from object (SPG) to scene (FS-COCO): the across-model mean shift is essentially zero ($\Delta = +0.004$, $|\Delta| = 0.034$ in fractional depth), so the same layers carry grounding signal across domains. The same heads are selected in both domains, not only the same depths: on average the five models share 5.4 of their top-10 heads, against fewer than one expected by chance, and selection frequencies correlate at $r = +0.74$ (App. E). Along the pilot's abstraction axis, by contrast, the peak moves earlier (Fig. 2), so the band is conserved across domains, not across levels of abstraction. The family split is the dominant axis of variation, not the input domain. Fig. 6 (bottom-left) characterises the signal carried by these heads: their combined-

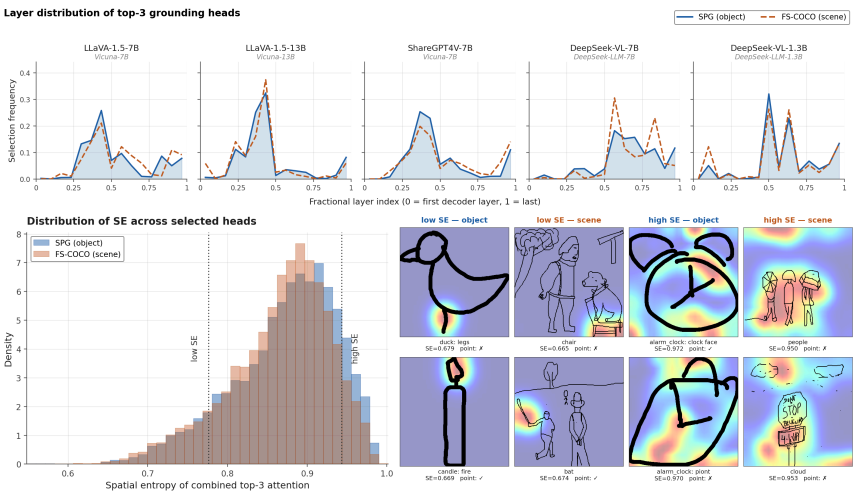

Figure 6: **Mechanistic profile of LVLM grounding. Top**: layer distribution of the top-3 grounding heads across the five models, SPG (object) vs FS-COCO (scene). **Bottom-left**: distribution of combined top-3 spatial entropy, pooled across models and datasets, with low/high SE tails marked. **Bottom-right**: LLaVA-1.5-13B attention overlays on samples drawn from each SE tail.

top-3 spatial entropy concentrates between 0.78 and 0.95, with overlapping SPG and FS-COCO distributions, indicating that domain shifts the band's content as little as it shifts its location. The qualitative samples in Fig. 6 (bottom-right) show what the SE tails correspond to spatially: low-SE samples tend to produce a single tightly-localised peak, whereas high-SE samples spread attention diffusely across the sketch with a correspondingly higher rate of pointing misses. SE thus measures whether the model produced a confident peak, while pointing accuracy measures whether that peak was correctly placed, and the two are complementary diagnostics rather than a single accuracy signal. Together, the three readouts portray grounding as a property of a stable, narrow layer band whose location follows the base language model in the two families we examine, while its signal quality varies with input.

Per-class temporal-spatial coherence. Tab. 1 and 2 both evaluate on SPG but probe different competences. We ask whether classes where the model agrees with humans on part-construction order are also classes the model grounds well, i.e., whether internalised convention (temporal) and localisation (spatial) are coupled at the category level. For each LLaVA variant we plot per-class LCS against per-class mAcc@0.1 across all twenty-five SPG categories (Fig. 8). The two probes dissociate in LLaVA-1.5-7B ($r = -0.13$, $\rho = -0.16$, n.s.) but couple in LLaVA-1.5-13B ($r = +0.40$, $\rho = +0.42$, $p < 0.05$): canonical-consensus categories (*suitcase*, *airplane*, *coffee_cup*, *butterfly*) cluster top-right, whilst ambiguous categories (*flower*) and outliers *drill* (high temporal, low spatial) and *candle* (mid temporal, low spatial) occupy the off-diagonal. LLaVA-1.5-7B orders parts at the level of a random permutation, so there is no ordering signal to couple with, whereas in LLaVA-1.5-13B the readouts align class by class, a diagnostic neither probe alone provides.

4.3 The layer band is predictive, not merely descriptive

Layer band as a head-selection prior. We test whether the mid-network band of Fig. 6 is predictive, not just descriptive. Restricting head selection to each family's empirical band ($[0.35, 0.45]$ Vicuna, $[0.50, 0.65]$ DeepSeek) matches or exceeds full-grid scoring in every (model, dataset) cell, with the largest gains on smaller models (LLaVA-1.5-7B: $+0.08$/$+0.09$ mAcc@0.1 on SPG/FS-COCO). High-scoring heads outside the modal range generalise less reliably, and the band acts as a noise-reducing prior, helping most where the model is weakest.

Model	SPG [34] Δ (full → band)	FS-COCO [19] Δ (full → band)
LLaVA-1.5-7B	+0.08 (0.65 → **0.73**)	+0.09 (0.69 → **0.78**)
LLaVA-1.5-13B	+0.01 (0.72 → **0.73**)	+0.03 (0.78 → **0.81**)
ShareGPT4V-7B	+0.04 (0.73 → **0.76**)	+0.05 (0.73 → **0.78**)
DeepSeek-VL-1.3B	+0.04 (0.56 → **0.60**)	+0.04 (0.61 → **0.65**)
DeepSeek-VL-7B	+0.01 (0.48 → **0.49**)	+0.04 (0.65 → **0.69**)

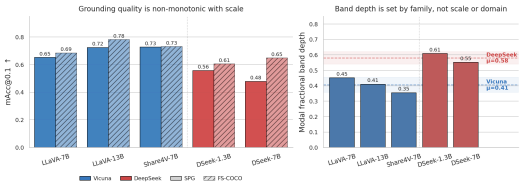

Figure 7: **Parameter count vs model family.** Left: mAcc@0.1 per (model, dataset). Right: modal fractional band depth per model, with family means (dashed) ± SD (shaded). Quality moves with scale, but band location does not.

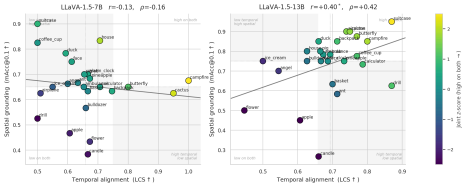

Figure 8: **Per-class temporal-spatial coherence on SPG.** LCS vs mAcc@0.1 across 25 categories, coloured by joint z-score. Probes dissociate in LLaVA-1.5-7B ($r = -0.13$) and couple in LLaVA-1.5-13B ($r = +0.40$, $p < 0.05$).

The selected heads carry the signal. Head selection is correlational, so we intervene, zeroing head outputs in the forward pass of LLaVA-1.5-13B. To avoid reading grounding off the ablated heads, we measure a downstream behaviour: whether the model can say on which side of the sketch a named part lies. Ablating the three most frequently selected heads lowers the logit margin towards the correct side by 0.118 (95% CI 0.074 to 0.163, $p = 0.0001$, clustered over 24 categories), against 0.001 for three random heads and 0.010 for three others in the same layers. Side accuracy barely moves (App. F), but the layer-matched control ties the effect to the selected heads, not to their depth.

Within- and across-family scaling. Across the five models, scale and grounding quality interact non-monotonically (Fig. 7, left). The Vicuna pair improves with size (LLaVA-1.5-7B $\rightarrow$ 13B: $+0.037$/$+0.084$ Point and $+0.072$/$+0.096$ mAcc@0.1 on SPG/FS-COCO), yet the DeepSeek pair reverses on objects (-0.077/-0.078 on SPG), with the smaller variant outperforming the larger. Crucially, scale and domain leave the modal grounding layer untouched (Fig. 7, right): within-family SD is 0.049 for the Vicuna trio and 0.040 for the DeepSeek pair, whilst the cross-family gap is 0.175, i.e., a $\sim$3.5$\times$ ratio. Per-model modal layers are identical to three decimal places across SPG and FS-COCO. In the two families we examine, the grounding band is therefore a property of the base LM family, not of parameter count or input domain, extending the split of Fig. 6 to individual models.

5 Discussion

What the readouts show that a score does not. Three observations here are invisible to an aggregate benchmark. First, model differences are structured: DeepSeek-VL-7B trails DeepSeek-VL-1.3B on object grounding, yet both place their grounding heads in the same band (Sec. 4.3). Second, failure is graded: under abstraction the model becomes wrong about *where* a part lies before its attention becomes diffuse (Sec. 3.1). Third, models defer to human convention: agreement rises and falls with human consensus, so variability in ambiguous categories is informative, not simply a lower score.

Limitations. The probes assume the queried part is present. Asked about absent parts taken from other categories, the model never signals absence and answers as it does for real parts, as in *"The bell of the airplane is located at the top of the wing"* and *"The stamens of the alarm clock are located at the top of the clock"* (App. G). The readout thus says *where* a named part is, not *whether* it is there, and pipelines in which parts may be absent need a separate presence check. The attention readouts also cover only two base language model families, and further families are future work.

6 Conclusion

We used sketches as a controlled abstraction test for LVLMs: strip texture and context, keep human-selected structure, and ask what survives. The resulting training-free probe yields three readouts: part-construction order tracks human convention where stable and is carried by the language model, frozen attention heads recover part regions without grounding supervision, and a narrow mid-network band set by the base language model in the families we examine acts as a head-selection prior that improves grounding without retraining and couples with the temporal readout at scale. Sketches thus become a diagnostic input rather than a domain-shift benchmark, exposing what LVLMs retain, when it becomes unreliable, and which layers carry it.

Acknowledgements

We acknowledge the support of UK Research and Innovation (UKRI) through the UKRI AI: Centre for Doctoral Training in AI for Digital Media Inclusion under Grant No.: EP/Y030915/1. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

References

[1] Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[2] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[3] Anthropic. Claude Haiku 4.5. https://www.anthropic.com/claude/haiku, October 2025. Released October 15, 2025; we use `claude-haiku-4-5`.

[4] Ellie Arar, Yarden Frenkel, Daniel Cohen-Or, Ariel Shamir, and Yael Vinker. SwiftSketch: A Diffusion Model for Image-to-Vector Sketch Generation. In *Proceedings of the SIGGRAPH Conference Papers*, 2025.

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A Frontier Large Vision-Language Model with Versatile Abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[6] Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[7] 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*, 2025.

[8] Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketching without Worrying: Noise-Tolerant Sketch-Based Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[9] Ayan Kumar Bhunia, Subhadeep Koley, Amandeep Kumar, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch2Saliency: Learning to Detect Salient Objects from Human Drawings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[10] Hakan Bilen and Andrea Vedaldi. Weakly Supervised Deep Detection Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[11] Ahmed Bourouis, Judith Ellen Fan, and Yulia Gryaditskaya. Open Vocabulary Semantic Scene Sketch Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[12] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding Everything: Emerging Localization Properties in Vision-Language Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[13] Hila Chefer, Shir Gur, and Lior Wolf. Transformer Interpretability beyond Attention Visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[14] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving Large Multi-modal Models with Better Captions. In *Proceedings of the European Conference on Computer Vision*, 2024.

[15] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect What You Can: Detecting and Representing Objects using Holistic Models and Body Parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[16] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[17] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. SceneTrilogy: On Human Scene-Sketch and its Complementarity with Photo and Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[18] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[19] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[20] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*, 2025. We use `gemini-2.5-flash-lite`.

[21] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[23] Mathias Eitz, James Hays, and Marc Alexa. How Do Humans Sketch Objects? *ACM Transactions on Graphics*, 2012.

[24] Judith E. Fan, Wilma A. Bainbridge, Rebecca Chamberlain, and Jeffrey D. Wammes. Drawing as A Versatile Cognitive Tool. *Nature Reviews Psychology*, 2023.

[25] Judith E. Fan, Robert X. D. Hawkins, Mike Wu, and Noah D. Goodman. Pragmatic Inference and Visual Abstraction Enable Contextual Flexibility during Visual Communication. *Computational Brain & Behavior*, 2020.

[26] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*, 2023.

[27] Songwei Ge, Vedanuj Goswami, C. Lawrence Zitnick, and Devi Parikh. Creative Sketch Generation. In *International Conference on Learning Representations (ICLR)*, 2021.

[28] David Ha and Douglas Eck. A Neural Representation of Sketch Drawings. In *International Conference on Learning Representations*, 2018.

[29] Daniel S. Hirschberg. A Linear Space Algorithm for Computing Maximal Common Subsequences. *Communications of the ACM*, 1975.

[30] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your Large Vision-Language Model Only Needs A Few Attention Heads For Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[31] Maurice G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 1938.

[32] Vladimir I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 1966.

[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[34] Ke Li, Kaiyue Pang, Jifei Song, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Honggang Zhang. Universal Sketch Perceptual Grouping. In *Proceedings of the European Conference on Computer Vision*, 2018.

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, 2024.

[37] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-Around Player? In *Proceedings of the European Conference on Computer Vision*, 2024.

[38] Bria Long, Judith Fan, Holly Huey, Zixian Chai, and Michael Frank. Parallel Developmental Changes in Children's Production and Recognition of Line Drawings of Visual Concepts. *Nature Communications*, 2024.

[39] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. DeepSeek-VL: Towards Real-World Vision-Language Understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[40] Timo Lüddecke and Alexander Ecker. Image Segmentation Using Text and Image Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[41] Kushin Mukherjee, Holly Huey, Xuanchen Lu, Yael Vinker, Rio Aguina-Kang, Ariel Shamir, and Judith E. Fan. SEVA: Leveraging Sketches to Evaluate Alignment between Human and Machine Visual Abstraction. In *Advances in Neural Information Processing Systems*, 2023.

[42] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards Interpreting Visual Information Processing in Vision-Language Models. In *International Conference on Learning Representations*, 2025.

[43] OpenAI. GPT-5 System Card. https://openai.com/index/gpt-5-system-card/, 2025. Released August 7, 2025; we use `gpt-5-nano`.

[44] 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*, 2024.

[45] Anran Qi, Yulia Gryaditskaya, Tao Xiang, and Yi-Zhe Song. One Sketch for All: One-Shot Personalized Sketch Segmentation. *IEEE Transactions on Image Processing*, 2022.

[46] Zhiyu Qu, Yulia Gryaditskaya, Ke Li, Kaiyue Pang, Tao Xiang, and Yi-Zhe Song. SketchXAI: A First Look at Explainability for Human Sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[48] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji Mullappilly, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel Grounding Large Multimodal Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[49] Leo Sampaio Ferraz Ribeiro, Tu Bui, John Collomosse, and Moacir Ponti. SketchFormer: Transformer-Based Representation for Sketched Structure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[50] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and Human Fixations: State-of-the-art and Study of Comparison Metrics. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[51] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. CLIP for All Things Zero-Shot Sketch-Based Image Retrieval, Fine-Grained or Not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[52] Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are Vision-Language Transformers Learning Multimodal Representations? A Probing Perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[53] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[54] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What Does CLIP Know About a Red Circle? Visual Prompt Engineering for VLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[55] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[56] Yael Vinker, Yuval Alaluf, Daniel Cohen-Or, and Ariel Shamir. CLIPascene: Scene Sketching with Different Types and Levels of Abstraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[57] Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. CLIPasso: Semantically-Aware Object Sketching. *ACM Transactions on Graphics (TOG)*, 2022.

[58] Qiang Wang, Haoge Deng, Yonggang Qi, Da Li, and Yi-Zhe Song. SketchKnitter: Vectorized Sketch Generation with Diffusion Models. In *International Conference on Learning Representations*, 2023.

[59] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic Segmentation Emerges from Text Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[60] 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*, 2024.

[61] Runpeng Yu, Weihao Yu, and Xinchao Wang. Attention Prompting on Image for Large Vision-Language Models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[62] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *Proceedings of the European Conference on Computer Vision*, 2014.

[63] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-Down Neural Attention by Excitation Backprop. In *Proceedings of the European Conference on Computer Vision*, 2016.

[64] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Appendix

Appendix Contents

A Implementation details **16**

A.1 Temporal probe prompt 16

A.2 Grounding pipeline 16

A.3 Compute 16

B Human consensus and LVLM agreement **16**

C Image controls for the temporal probe **17**

D Grounding at stricter overlap **17**

E Object and scene head overlap **18**

F Head ablation **18**

G Phantom parts **19**

H Newer open models **19**

A Implementation details

A.1 Temporal probe prompt

The temporal probe uses the system and user prompt below, reproduced verbatim. The part names are inserted in a shuffled order that is fixed per sketch, as a quoted, comma-separated list in set notation, so the listing carries no ordering cue. The response is parsed for the `order` field, and each returned name is matched to the given parts after lower-casing and normalising underscores and hyphens.

```
System: You analyse hand-drawn sketches and infer the temporal order in which a
human drew their parts. Your inference is based purely on visual evidence in the
sketch - relative size, position, layering, and which parts visually rest on or
extend from others - combined with general knowledge of how humans typically
construct drawings.

User: You are looking at a black-line sketch of a {class}.

The sketch contains the following parts (presented as an unordered set; the
listing order has no relation to the drawing order):
$\{"{part}", "{part}", ...\}$

Question: in what order did the human draw these parts?

Output a single JSON object with two fields:
- "reason": one sentence describing the visual evidence you used.
- "order": the parts arranged from earliest drawn to latest drawn.

Constraints:
- Use only parts from the set; do not invent or modify part names.
- Use each part exactly once.

Example format:
{"reason": "...", "order": ["first", "second", ...]}
```

A.2 Grounding pipeline

For each query we read the attention from the last input token to the image patches in every head (Sec. 3.3). Heads are scored by image-attention mass and spatial entropy, the candidates are cut with the chord-elbow rule of Kang et al. [30], and the top $k = 3$ are kept. Each selected map is mean-subtracted, passed through a ReLU, smoothed with a Gaussian of $\sigma = 1$, and normalised. The three maps are combined by element-wise maximum and binarised at the 80th percentile, so the predicted mask always covers the top 20% of the patch grid (24×24 for the LLaVA-1.5 models). The baselines binarise differently: CLIPSeg [40] thresholds its sigmoid output at 0.5, and GroupViT [59] assigns each pixel to its highest-scoring group. A fixed-area mask is favoured at a loose IoU threshold and penalised at a strict one, which App. D examines. SPG ground truth for a part is the convex hull of its strokes, and FS-COCO ground truth is the object mask from FS-COCO-Seg [11].

A.3 Compute

All experiments are forward passes of frozen models on a single NVIDIA A100 GPU. Closed-source models are queried through their vendor APIs. No model is trained or fine-tuned.

B Human consensus and LVLM agreement

Human consensus for a category measures how consistently SPG drawings follow its modal part order, as ranked in Fig. 4. Tab. 5 correlates it with each model's per-category LCS over the 25 categories. Memorisation and retained convention make different predictions here. A model reproducing stored SPG trajectories would agree equally well whether or not humans converge, whereas a model that has retained the convention should agree more where the convention is stronger. Eight of the nine models show a significant positive correlation. In those eight, agreement on high-consensus categories exceeds that on low-consensus categories by 0.09 to 0.24 LCS. The exception is LLaVA-1.5-7B, whose agreement is at the level a random permutation of the same parts would achieve, so there is no ordering signal for consensus to track.

An independent check uses the model's own consensus, the agreement among its predictions across the sketches of a category. It also correlates with human consensus, in all nine models ($r = +0.35$ to $+0.79$), significantly in eight. With 25 categories both analyses are modestly powered, and we read them as evidence rather than proof.

Table 5: **Human consensus against LVLM agreement.** Correlation between per-category human consensus (Fig. 4) and per-category LCS with the human trajectory, over the 25 SPG categories.

Model	**Pearson** r	p	**Spearman** ρ
Gemini 2.5 Flash-Lite [20]	$+0.82$	< 0.001	$+0.77$
Claude Haiku 4.5 [3]	$+0.77$	< 0.001	$+0.80$
Qwen2.5-VL-7B [7]	$+0.71$	< 0.001	$+0.66$
Molmo-7B-D [21]	$+0.70$	< 0.001	$+0.64$
GPT-5 (nano) [43]	$+0.61$	0.001	$+0.64$
InternVL2-8B [16]	$+0.60$	0.002	$+0.59$
LLaVA-1.5-13B [35]	$+0.54$	0.005	$+0.59$
Qwen2.5-VL-3B [7]	$+0.54$	0.006	$+0.48$
LLaVA-1.5-7B [35]	-0.06	n.s.	-0.25
Mean over nine models	$+0.58$		

C Image controls for the temporal probe

Protocol. The controls use 50 sketches per category and the prompt of App. A.1. Each sketch has one shuffle of its parts, drawn once with a fixed seed and reused in every condition, and the object name in the prompt is always the true category. *Blank canvas* replaces the sketch with a 512×512 white image. *No image* sends the prompt alone, with no image tokens. *Mismatched category* shows a sketch from the category whose part vocabulary overlaps least with the target's. Parts that a response omits are appended in their shuffled input order, so a response that cannot be parsed scores at the level of a random ordering rather than of the reference. Each control is compared with the real sketch on per-category means by a Wilcoxon signed-rank test over the 25 categories. Two models, three controls, and three metrics give the eighteen comparisons of Sec. 4.2. None favours the real sketch, and the only significant one favours the no-image condition for Qwen2.5-VL-7B on Pair ($p = 0.04$).

Deviation set. The main protocol samples sketches whose drawing order matches the category's modal order, so every sketch in a category has the same correct answer, which the text alone can supply. The deviation set instead takes sketches with the same parts as the modal order but a different drawing order. The prompt is then identical to the main protocol and only the image, and with it the correct answer, changes. We test whether the prediction is closer to the sketch's own order than the modal order is. A per-sketch shift in that direction does not survive clustering by category ($p = 0.26$, 17 categories).

Forced choice. The model sees one sketch and two candidate orderings, its own and another, matched for edit distance from the modal order, and must say which belongs to the sketch. A model that reasons from drawing conventions alone cannot prefer either, so chance is 50% by construction. Qwen2.5-VL-7B is at chance both with the sketch and without it.

D Grounding at stricter overlap

Tab. 6 re-scores every method at three IoU thresholds from the stored per-part IoU, with no new inference. Two mAcc@0.1 entries differ from Tab. 2 by 0.01, and we report this run unchanged so that no row mixes runs. Grounding is evaluated on the same sketches for every method, so we pair the comparison by category with a Wilcoxon signed-rank test (25 categories for SPG, 185 for FS-COCO). At mAcc@0.1 LLaVA-1.5-13B is above both trained grounders on both datasets, and all four comparisons are significant ($p < 0.002$). At mAcc@0.3 it remains above both, significantly in three of four comparisons. At mAcc@0.5 CLIPSeg is nominally ahead on both datasets and GroupViT on SPG, and no difference is significant.

Table 6: **Grounding at three overlap thresholds.** All methods re-scored from one run.

	SPG [34]			FS-COCO [19]		
Model	**@0.1↑**	**@0.3↑**	**@0.5↑**	**@0.1↑**	**@0.3↑**	**@0.5↑**
CLIPSeg [40]	0.27	0.17	0.12	0.40	0.31	0.20
GroupViT [59]	0.58	0.28	0.11	0.60	0.21	0.08
LLaVA-1.5-13B [35]	0.72	0.29	0.06	0.78	0.35	0.11

Threshold sweep. To locate the cause of the drop at strict overlap, we sweep the binarisation percentile over $\{60, 70, 75, 80, 85, 90, 95\}$ for LLaVA-1.5-13B on SPG. At the 80th percentile used throughout, the mask has precision 0.35 and recall 0.54 and covers about 1.5 times the ground-truth area. mAcc@0.5 rises as the percentile is lowered, that is, as the mask grows. The shortfall at strict overlap therefore comes from under-coverage at a fixed threshold rather than from an over-large mask. Two properties of the evaluation bear on all thresholds. The LVLM mask has a fixed area while CLIPSeg's adapts to the part (App. A.2), and the convex-hull ground truth of SPG is larger and smoother than the inked region, which favours diffuse predictions.

E Object and scene head overlap

For each model and dataset we count how often each (layer, head) cell enters the top-3 selection across all grounding queries. Tab. 7 compares SPG (object) with FS-COCO (scene) by the modal fractional depth of these counts, by the number of the ten most frequently selected heads that the two datasets share, and by the Pearson correlation of selection frequencies over all cells. The datasets have different queries, so the comparison is between distributions rather than paired queries. Two unrelated top-10 sets drawn from a grid of $L \times H$ cells share $100/(LH)$ heads in expectation, well under one for every model here. The modal depth is identical to three decimal places in every model, and on average 5.4 of the top ten heads are shared.

Table 7: **Grounding-head overlap between object and scene sketches.** Depth is the modal fractional depth of top-3 selections.

Model	**Depth, SPG**	**Depth, FS-COCO**	**Shared top-10**	**Frequency r**
ShareGPT4V-7B [14]	0.355	0.355	6/10	+0.72
LLaVA-1.5-13B [35]	0.410	0.410	5/10	+0.74
LLaVA-1.5-7B [35]	0.452	0.452	5/10	+0.66
DeepSeek-VL-7B [39]	0.552	0.552	4/10	+0.70
DeepSeek-VL-1.3B [39]	0.609	0.609	7/10	+0.87
Mean			5.4/10	+0.74

F Head ablation

We ablate a head by zeroing its slice of the input to the attention output projection, which removes its contribution to the residual stream, in every forward pass of LLaVA-1.5-13B. Each query asks *"Is the {part} of the {category} on the left or the right side of this sketch? Answer with one word."* The correct answer is the side on which the centroid of the part's ground-truth mask falls, and parts whose centroid lies within 0.12 of the image width from the midline are excluded, so a correct answer requires localisation. We read the next-token logits rather than generating, take the maximum logit over the tokens for *left* and for *right* (with and without capitalisation and a leading space), and define the margin as the correct side's logit minus the other's.

The top-3 heads are the three most frequently selected across the SPG grounding queries of LLaVA-1.5-13B. Random heads are drawn uniformly from the rest of the grid, and layer-matched heads from the other heads in the layers of the top-3, both with a fixed seed. Per-category mean margins are compared with the unablated model by a Wilcoxon signed-rank test over the 24 categories with queries in every condition, with a 95% bootstrap interval over categories. Paired over categories, the top-3 drop also exceeds the random ($p = 0.0002$) and layer-matched ($p = 0.0003$) drops. The

majority-class rate for side accuracy is 0.586. Accuracy barely moves under any ablation (Tab. 8), since three of 1,600 heads are too few to flip many decisions, while the margin shows that the selected heads, and not other heads at the same depth, carry the signal.

Table 8: **Head ablation in LLaVA-1.5-13B.** Side accuracy and logit margin towards the correct side.

Ablated heads	Side accuracy↑	Margin↑	Change in margin
None	0.701	0.854	
Top-3 grounding heads	0.713	0.736	$\mathbf{-0.118}$ ($p = 0.0001$)
Random 3 heads	0.701	0.854	-0.001
3 heads, same layers	0.695	0.844	-0.010

G Phantom parts

To test what the probe does when there is nothing to find, we query parts that are absent from the sketch. For each SPG category we draw two part names from the vocabularies of other categories, excluding any name that occurs in the target category under any spelling, and query them on ten sketches per category with the spatial template of Sec. 3.3, alongside the real parts of the same sketches. We also pose the same question for a text answer. The model never states that a part is absent, and its answers about phantom parts read like its answers about real ones:

- phantom: *"The bell of the airplane is located at the top of the wing."*
- phantom: *"The stamens of the alarm clock are located at the top of the clock."*
- real: *"The body of the airplane is located in the middle of the drawing."*

The probe therefore localises without verifying existence. It answers where a named part is, not whether it is there, and a pipeline in which parts may be absent needs a separate presence check.

H Newer open models

The temporal probe needs only generated text, so it extends directly to newer releases. We ran Qwen3-VL [6] at 4B and 8B under the identical protocol, which gives a generational comparison against the Qwen2.5-VL rows of Tab. 1 with family and probe held fixed (Tab. 9). The generational gain is large at the smaller scale ($+0.16$ Edit and $+0.20$ Pair from Qwen2.5-VL-3B to Qwen3-VL-4B) and small at the larger ($+0.03$ and $+0.03$ from 7B to 8B). The two Qwen3-VL sizes are almost indistinguishable, so agreement has saturated within the family. Qwen3-VL-4B matches or exceeds GPT-5 (nano) and Gemini 2.5 Flash-Lite on all three metrics, and the gap to Claude Haiku 4.5 narrows to 0.04 LCS. In line with App. C, these numbers measure the drawing convention the language model has retained.

Table 9: **Part-construction order for Qwen3-VL** on SPG, protocol as in Tab. 1.

Model	Edit↑	LCS↑	Pair↑
Qwen2.5-VL-3B (Tab. 1)	0.57	0.77	0.64
Qwen3-VL-4B	**0.73**	**0.85**	**0.84**
Qwen2.5-VL-7B (Tab. 1)	0.69	0.84	0.80
Qwen3-VL-8B	**0.72**	**0.85**	**0.83**
Claude Haiku 4.5 (Tab. 1)	0.80	0.89	0.88

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three findings: part-construction order tracks human convention and is carried by the language model, attention heads recover part regions without grounding supervision, and the grounding signal sits in a mid-network band that follows the base language model in the two families examined. Sections 4.2 and 4.3 and Appendices B to H support them, and the abstract, Section 1 and Section 4.1 state which models each probe covers.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 contains a Limitations paragraph. The probes assume the queried part is present, and the model never signals an absent part (Appendix G). The attention readouts cover two base language model families, so the link between band position and lineage is established for those families only. Section 4.2 and Appendix C show that the temporal ordering is carried by the language model rather than recovered from the image, and Section 4.2 and Appendix D report that the advantage over trained grounders holds at coarse and moderate overlap but not at strict overlap.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results requiring proofs. All claims are empirical.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 describes the three probes, including prompt templates, head-selection criteria (image-attention mass, spatial entropy, chord-elbow rule), the aggregation pipeline (Gaussian smoothing, element-wise maximum, percentile thresholding) and the evaluation metrics. Section 4.1 specifies the model sets, decoding settings and sampling protocol (100 strict-mode-matching sketches per SPG category). Appendix A gives the verbatim temporal prompt and the grounding hyperparameters, and Appendices B to H give the protocol of every additional analysis. All probes use frozen pre-trained models with no fine-tuning.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and per-sample outputs are not released with the paper, and we plan to release the probing pipeline and per-sample outputs. SPG, FS-COCO, FS-COCO-Seg and Pascal-Part are publicly available, and all evaluated LVLMs are available through HuggingFace or vendor APIs.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 and Appendix A specify model variants, deterministic decoding (temperature 0), sample sizes per category, prompt templates, head selection (chord-elbow rule, top-$k = 3$), Gaussian smoothing ($\sigma = 1$), the 80th-percentile mask threshold and the evaluation metrics. The per-family layer bands used in Section 4.3 are given explicitly.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 reports mean $\pm$ SD across the 25 SPG categories (each averaged over 100 strict-mode-matching sketches). The image controls (Section 4.2, Appendix C), the stricter grounding thresholds (Appendix D) and the head ablation (Section 4.3, Appendix F) use Wilcoxon signed-rank tests paired over categories, and the ablation also reports a 95% bootstrap confidence interval over categories. The correlations with human consensus (Appendix B) and the per-class temporal-spatial correlations (Section 4.2) report Pearson and Spearman coefficients with p-values.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix A.3 states that all experiments are forward passes of frozen pretrained models on a single NVIDIA A100 GPU and that closed-source models are queried through vendor APIs. Run times and memory use are not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research uses publicly available datasets and pre-trained models, conducts no human-subjects experimentation, and does not introduce new data or models with foreseeable misuse risk.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work analyses pre-trained LVLMs for interpretability and does not introduce new generative capabilities. Section 1 discusses positive uses, such as reading region grounding directly from an LVLM in sketch-based retrieval, segmentation and editing pipelines without fine-tuning. Section 5 notes the main risk in use: the readout never signals an absent part, so a pipeline in which parts may be absent needs a separate presence check. We see no other negative societal impact distinct from those of the underlying LVLMs.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: The paper does not release new pre-trained models, datasets, or generative tools. It analyses existing, publicly released LVLMs.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets (SPG, FS-COCO, FS-COCO-Seg, Pascal-Part), the ControlSketch abstraction method and all models (LLaVA-1.5, ShareGPT4V, DeepSeek-VL, Qwen2.5-VL, Qwen3-VL, Molmo, InternVL2, Claude Haiku 4.5, GPT-5 nano, Gemini 2.5 Flash-Lite) are cited with their original publications, as are the CLIPSeg and GroupViT grounders.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new datasets, models, or pre-trained checkpoints. The analysis pipeline is built on existing components (the Localization Heads recipe and classical sequence metrics).

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects research. Human part-construction trajectories used as ground truth are taken from the publicly released SPG dataset, which provides per-stroke part annotations recorded during the dataset's original construction.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: As stated in question 14, this paper does not involve human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs (specifically LVLMs) are the subject of analysis in this work, not a tool used in core method development. Pre-trained LVLMs are evaluated as black boxes without fine-tuning, training, or modification of their parameters. LLMs were used for routine writing assistance (editing, formatting) which does not require declaration per NeurIPS policy.

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.