VLMs Need Words: Vision Language Models Ignore Visual Detail In Favor of Semantic Anchors

Haz Sameen Shahgir[1], **Xiaofu Chen**[2], **Yu Fu**[1], **Erfan Shayegani** [1]
Nael Abu-Ghazaleh[1], **Yova Kementchedjhieva**[2], **Yue Dong**[1]

University of California, Riverside [1], MBZUAI[2]

hshah057@ucr.edu, yued@ucr.edu

Abstract

Vision-language models (VLMs) have achieved impressive performance across a wide range of multimodal tasks. However, they often fail on tasks that require fine-grained visual perception, even when the required information is still present in their internal representations. Prior work has attributed this "hidden-in-plain-sight" gap to the language model, but the cause remains unexplained. In this work, we demonstrate that this gap arises from the language model's lack of semantic labels for fine-grained visual details: when visual entities can be mapped to known concepts, VLMs bypass visual comparison and reason through language; when they cannot, VLMs resort to brittle and hallucinated descriptions. We verify this across semantic correspondence, synthetic shape matching, and face matching, and find that VLMs perform much better when the relevant entities are nameable than when they are unnamable. Mechanistically, Logit Lens analysis confirms that VLMs explicitly recover semantic labels for nameable entities and surface more unique tokens compared to unnameable entities. Furthermore, we show that this limitation can be addressed: teaching completely arbitrary names for unknown entities improves performance. More importantly, task-specific finetuning yields even stronger generalization without relying on language priors, i.e. through real visual perception. Our findings suggest that current VLM failures on visual tasks reflect a learned shortcut rather than a fundamental limitation of multimodal reasoning[1].

1 Introduction

Vision Language Models (VLMs) have achieved impressive performance across a wide range of multimodal tasks, from visual question answering and image captioning (OpenAI, 2024) to document understanding and visual grounding (Bai et al., 2023). These models are pre-trained on billions of image-text pairs and further finetuned on curated instruction-following datasets, with the implicit promise that this pipeline produces general-purpose multimodal reasoners capable of handling novel multimodal problems out of the box.

Yet this promise remains unfulfilled. VLMs consistently fail on tasks that demand fine-grained visual perception. They struggle with chart and diagram understanding (Zhang et al., 2024), misinterpret optical illusions that humans resolve effortlessly (Shahgir et al., 2024), falter on abstract visual reasoning (Chollet et al., 2026), and most strikingly, cannot solve basic visual discrimination tasks that three-year-old children handle with ease (Chen et al., 2026). These failures are not domain- or model-specific; they emerge whenever models must reason from pixels instead of matching patterns. These failures are often mitigated through task-specific supervision, but such improvements do not explain why VLMs pretrained on web-scale data fail to generalize in the first place. We argue that the failure is rooted in the standard pretraining-and-SFT pipeline itself, which encourages models to rely on semantic shortcuts rather than learning transferable visual skills.

[1]Code and data are publicly available at github.com/Patchwork53/VLMs-Need-Words-COLM2026

Recently, Fu et al. (2025) and Liu et al. (2025) have shown that VLMs' internal representations retain enough visual information to solve many of these challenging tasks, even when the model's verbal output fails. This "hidden-in-plain-sight" gap was identified as a shortcoming of the language model (LM) backbone, but its underlying mechanism was left unexplored. At the same time, other work has documented a systematic over-reliance on language priors: VLMs underperform on tasks that resist mediation through language (Fu et al., 2024) and exhibit biases inherited from their LM backbone (Vo et al., 2026).

In this work, we connect and explain these observations. We hypothesize that VLMs short-circuit visual reasoning by mapping visual entities to discrete semantic labels in the language space. When a visual entity can be named, the model bypasses pixel-level comparison and transfers the task to the language space. When no crisp label exists, the model still attempts the same linguistic strategy, generating approximate, often hallucinated descriptions that actively degrade its reasoning. This explains both the hidden-in-plain-sight gap and why failures systematically cluster around novel or sub-semantic visual content.

We test this hypothesis through a visual correspondence task, where models must identify matching entities between two images, a capability that underlies real-world applications from medical differential diagnosis to temporal change detection in videos. Correspondence is a natural testbed because the degree to which target entities admit semantic labels varies naturally across settings: some keypoints on an object have well-known names (e.g., the *pedal* of a bicycle; see Figure 1, point D), while others do not (e.g., the joint between the head tube and top tube of a bicycle; see Figure 1, point A). We extend this contrast by constructing synthetic tasks that pair common shapes ("star," "circle") against procedurally generated ones, and recognizable celebrity faces against AI-generated unknown faces.

Across all three settings, we find the same pattern. VLMs perform substantially better on nameable entities than on unnameable ones, even when representation probing confirms that internal features contain sufficient information in both cases. Chain-of-thought reasoning disproportionately benefits named entities, consistent with the model leveraging verbal reasoning to recover and match discrete labels. Logit Lens analysis (nostalgebraist, 2020) confirms the mechanism: for known entities, hidden states progressively resolve from semantically unrelated tokens through approximate descriptors to exact labels (e.g., "triangle" to "pointed" to "star"), while unknown entities remain semantically indiscernible.

We further show that teaching VLMs completely arbitrary names for novel shapes substantially closes the performance gap. Once a shape has a name, the model can short-circuit visual comparison altogether, matching shapes by their learned labels rather than their pixels. The Logit Lens signature of the name-trained models confirms this observation: downstream accuracy grows proportionally to the increase in semantic discernibility.

However, semantic labeling is sufficient but not necessary for closing this gap. We investigate what task-specific finetuning changes in VLMs. When finetuned directly on the correspondence task, VLMs outperform name-trained models while exhibiting lower semantic discernibility. This indicates that task-specific finetuning teaches a distinct mechanism — direct visual comparison — that does not route through language. Together, these findings suggest that current VLM failures on visual tasks reflect a learned shortcut induced by the standard pretraining-and-SFT pipeline, not a fundamental limitation of the architecture.

1. We show, across semantic, shape, and face correspondence tasks, that VLM performance depends strongly on whether the target visual entity admits a semantic label, rather than on whether the internal representations contain sufficient information to solve the task.
2. We provide mechanistic evidence via Logit Lens analysis that known entities become progressively more semantically explicit inside the language model, while unknown entities remain indiscernible, and that teaching arbitrary names to unknown entities closes this discernibility gap.
3. We demonstrate that semantic labeling is sufficient but not necessary for closing the representation-to-output gap: teaching names enables a linguistic shortcut, while direct task finetuning elicits genuine visual reasoning rather than language-mediated semantic reasoning.

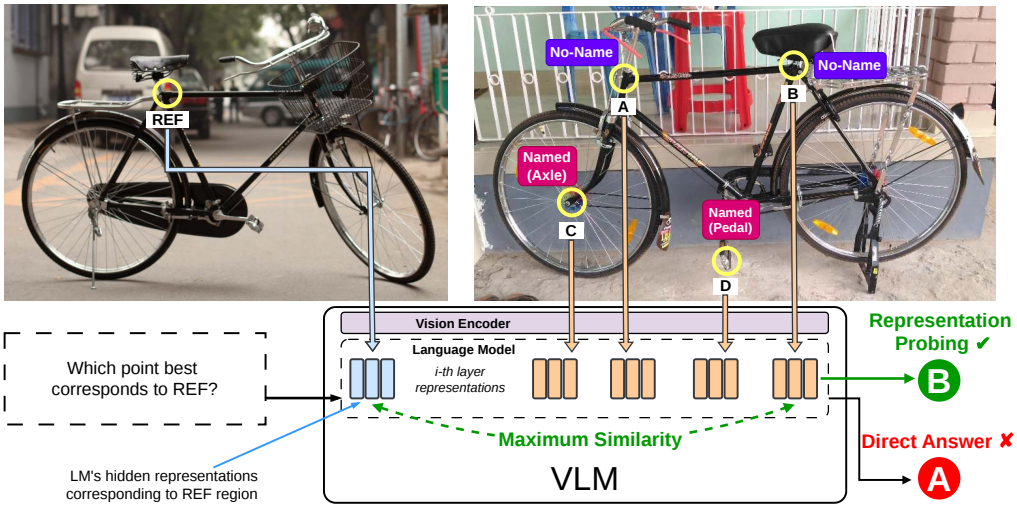


Figure 1: Overview of the correspondence task framed as multiple-choice VQA. The VLM must match the reference point "REF" in the first image to one of four candidate regions (A, B, C, D) in the second image. Regions can either be Named or have No-Name. We evaluate VLMs in three setups: Direct answer, Chain-of-Thought, and Representation Probing.

2 The Performance Gap Between Textual Response and Representation Space is Larger for Unknown Entities

In this section, we first study the role of semantic anchors by comparing VLM performance on semantic correspondence for named vs. unnamed entities in real images, then isolate this effect using synthetically generated 2D object and face correspondence tasks. Beyond fine-grained visual correspondence, we also study the effect of semantic anchors on Art Style Matching, where VLMs need to identify one of four image options that has the same art style as a target image. This task requires holistic understanding of

Experimental Setup For all correspondence tasks in this section, we follow the methodology of Fu et al. (2025) and add Chain-of-Thought (CoT) evaluation to study the effect of textual reasoning on inherently vision-heavy tasks. We frame the problem as a multiple-choice visual question answering (MC-VQA) task. We evaluate VLMs in three setups: 1) MC-VQA with Direct Answer, 2) MC-VQA with CoT, and 3) Representation Probing as shown in Figure 1. For MC-VQA, we present a VLM with two images and prompt it to find a point in the second image that matches a reference point in the first image. We annotate the reference point with "REF" and different regions in the second image with the four options ("A", "B", "C", and "D"). For Direct, we prepend *"The correct answer is ("* to the assistant response and directly generate the answer. This is the setup explored by Fu et al. (2025); Liu et al. (2025), which requires the VLM to complete the task entirely through non-verbal reasoning in a single forward pass. In our work, we also test Chain-of-Thought, which not only allows VLMs to reason through text but also allows them to attend to the visual tokens multiple times before generating the answer. We append *"Think step by step before choosing an option."* to the original prompt. We impose no output-format constraints on the VLM, thereby avoiding the diversion of model capacity from the correspondence task to format compliance. Instead, we use Qwen3-8B to extract the selected answer option from the VLM's free-form response. We use the default sampling parameters for all VLMs. In the Representation Probing setup, we define a square region of 30×30 pixels for each point of interest and extract the corresponding visual tokens. For each layer, we compute the similarity between the hidden representations of the "REF" point in the first image and those of the four options in the second image. A sample is considered correct if the ground-truth option has the highest similarity. Since each region maps to multiple tokens, we use the MaxSim operator (Khattab & Zaharia, 2020) instead of cosine similarity, which requires a lossy pooling step.

2.1 Testing Nameable and Unnameable Key Points in Semantic Correspondence

Fu et al. (2025) and Liu et al. (2025) have demonstrated that using VLMs' internal representations consistently outperforms their textual output on semantic correspondence on the SPairs71k (Min et al., 2019) dataset. SPair-71k consists of a total of 70,958 image pairs from PASCAL 3D+ (Xiang et al., 2014) and PASCAL VOC 2012 (Everingham et al., 2014) spanning 18 object categories encompassing vehicles (e.g., "aeroplane"), ordinary objects (e.g., "bottle"), and animals (e.g., "sheep").

We isolate the effect of nameability by benchmarking VLMs on named and unnamed splits of the SPair-71k test set images. SPair-71k provides consistent keypoint annotations across all images within a category. We manually categorize each keypoint as either *Named* or *No-Name*: keypoints corresponding to well-known parts (e.g., pedal, seat, handlebar) are *Named*, while keypoints at ambiguous locations (e.g., the junction between the handlebars and the stem) are *No-Name*. Each task is assigned to the split corresponding to its reference keypoint, resulting in 11,362 Named and 11,792 No-Name samples.

Model	Size	Subset	Direct (D)	CoT (C)	C−D Δ	Rep. Probe (R)	R−max(D,C) Δ
Qwen3VL	2B	Named	36.6	57.4	**20.8**	68.7	11.3
		No-Name	32.4	42.2	9.8	60.2	**18.0**
	4B	Named	52.5	62.9	**10.4**	68.4	5.5
		No-Name	37.2	46.7	9.5	62.1	**15.4**
	8B	Named	53.8	65.3	**11.5**	68.6	3.3
		No-Name	40.3	47.9	7.6	61.6	**13.7**
InternVL3.5	2B	Named	28.1	34.0	**5.9**	56.8	**22.8**
		No-Name	24.8	30.0	5.2	47.4	17.4
	8B	Named	34.3	46.7	**12.4**	56.7	10.0
		No-Name	28.2	34.4	6.2	51.1	**16.7**
	14B	Named	31.9	43.4	**11.5**	56.4	13.0
		No-Name	28.4	36.5	8.1	53.8	**17.3**

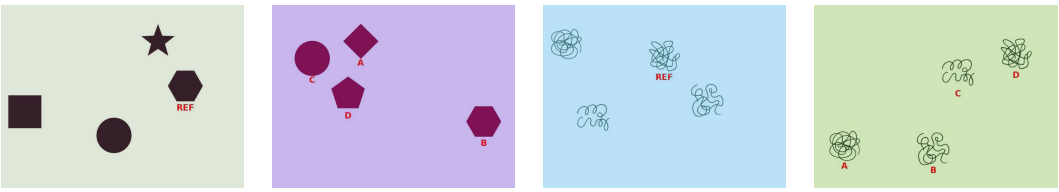
Table 1: Semantic correspondence results on SPairs71k. R denotes the representation-probing accuracy from the **best-performing layer** of the language model. We report layer-wise performance in Appendix Fig. 8. The R−max(D, C) Δ column shows the gap between Representation Probing and the best textual output.

Table 1 shows results across two VLM families on the Named and No-Name subsets of SPairs71k. First, the accuracy of VLMs on the No-Name subset is lower than on the Named subset for all model sizes, across both model families, and in both the Direct and CoT setups. This shows that VLMs struggle with the correspondence task when the reference point lacks a definitive semantic label. Second, we find that the gap between Representation Probing and best textual strategy (R−max(D, C) Δ) is consistently larger for the No-Name subset: Qwen3VL-8B achieves an R−max(D, C) gap of just 3.3% on Named keypoints, nearly closing the gap when semantic anchors are available, but 13.7% on No-Name, while unnamed entities remain bottlenecked. The pattern holds for InternVL3.5 at 8B and 14B. In contrast, InternVL3.5-2B appears to lack the baseline capability required for the task, achieving only 28.1% accuracy even on the Named subset, barely above the 25% random-chance baseline. We report the performance of Gemma-3-4B and 12B in Appendix E.

Chain-of-thought reasoning disproportionately benefits named entities: Qwen3VL-2B gains +20.8% from CoT on Named keypoints but only +9.8% on No-Name. Inspecting the CoT chain (See Appendix Figure 7), we find that VLMs explicitly generate the names of the points when available, effectively converting the visual correspondence task to a verbal string-matching task. This CoT advantage narrows with scale (Qwen3VL-8B: +11.5% Named vs. +7.6% No-Name), consistent with larger models performing more of this reasoning internally. Representation Probing performance, by contrast, remains stable across model sizes within each family (e.g., 68.4–68.7% Named and 60.2–62.1% No-Name for Qwen3VL). Since models within a family share similar vision encoders, this stability indicates that

the LLM layers do not destroy visual information; rather, larger LMs are more effective at surfacing it through text.

2.2 Confirming the Effect of Semantic Anchors on Shape and Face Correspondence



(a) Known Ref. (b) Known Tgt. (c) Unknown Ref. (d) Unknown Tgt.

Figure 2: Example of our 2D shape correspondence task. Face correspondence example is presented in Appendix Fig. 11.

The SPairs71k Semantic Correspondence task uses real-world images of entities subject to factors such as occlusion, varying resolution, varying entity size relative to the image, and visual artifacts. Since VLMs operate by converting non-overlapping image patches into tokens, the effects of these factors are complex and hard to control. To study the effect of semantic anchors in isolation, we design the 2D object correspondence and face correspondence tasks. This synthetic setup allows us to control resolution and entity size, and ensures there is no partial occlusion or visual artifacts. Figure 2 shows an example of the 2D object correspondence task. To construct the dataset, we use 10 common shapes ("square", "circle", "star", etc.) as known 2D objects and generate unknown 2D shapes ("squiggles") using a randomized spline-generation algorithm whose complexity can be controlled. Squiggles are generated by iteratively sampling n anchor points. A cubic spline is then fit through the resulting anchors and rendered with $4\times$ supersampling for anti-aliasing.

For the face correspondence task, we first collect synthetic faces from FluxSynID (Ismayilov et al., 2025) and a single photograph of well-known celebrities. We deliberately restrict the dataset to a single demographic group so that models cannot solve the task using coarse cues such as ethnicity or gender. We use East Asian male identities because our pilot experiments found that the evaluated VLMs recognized celebrities from this group more consistently than celebrities from other groups, such as African male celebrities. We then prompt `Nano-Banana-2` (Google DeepMind, 2025) with one example image to generate 4 images per person in different poses, lighting conditions, and environments. Because both the known and unknown celebrity images are generated through the same `Nano-Banana-2` pipeline, any model-specific visual artifacts are controlled for. We verify that the VLMs recognize each celebrity via knowledge probing and construct VLM-specific splits of the known celebrity dataset accordingly. The probing methodology and exact splits are reported in Appendix C. Our probing experiments reveal that InternVL3.5 is unable to recognize most celebrities. We therefore evaluate the Gemma3 model family to ensure our findings generalize across VLM architectures. We discuss InternVL3.5's inability to identify celebrities and its performance on Shape Correspondence in Appendix D.

Results Table 2 shows that in this controlled synthetic setting, the gap between known and unknown entities observed in Section 2 is amplified. The gap between Rep. Probing and the best textual baseline (R−max(D,C) Δ) is consistently larger for unknown entities across all models on both tasks. On 2D shapes, the gap is particularly stark: Qwen3VL-2B achieves just 29.0% Direct accuracy on unknown shapes versus 74.2% via Rep. Probing, a 45.2 percentage point gap, while known shapes are saturated (100% Rep. Probe, 97.3% CoT).

Unlike the semantic correspondence results, CoT can be harmful on these tasks: several models show negative C−D deltas on unknown entities (e.g., Qwen3VL-8B drops 19.4 points with CoT on unknown shapes), suggesting that without a semantic anchor, verbal reasoning devolves into hallucinated descriptions that actively mislead (see Appendix G.3). As before, Representation Probing performance remains stable across model sizes within each family, while larger models narrow the textual output gap.

Model		Known	Faces					2D Shapes				
			Direct (D)	CoT (C)	C−D Δ	Rep. Probe (R)	R−max(D,C) Δ	Direct (D)	CoT (C)	C−D Δ	Rep. Probe (R)	R−max(D,C) Δ
Qwen3VL	2B	✓	77.1	59.2	−17.9	97.2	20.1	54.1	97.3	43.2	100	2.7
		×	41.1	37.4	−3.7	70.8	**29.7**	29.0	27.3	−1.7	74.2	**45.2**
	4B	✓	84.2	83.2	−1.0	93.2	9.0	93.5	99.4	5.9	100	0.6
		×	56.9	55.7	−1.2	73.4	**16.5**	48.4	40.0	−8.4	91.7	**43.3**
	8B	✓	83.9	85.2	1.3	92.8	7.6	99.7	99.9	0.2	100	0.1
		×	65.7	63.8	−1.9	74.1	**8.4**	57.1	37.7	−19.4	86.1	**29.0**
Gemma3	4B	✓	49.8	52.7	2.9	61.2	8.5	50.6	70.5	19.9	98.9	28.4
		×	32.4	31.9	−0.5	42.9	**10.5**	30.5	32.3	1.8	91.7	**59.4**
	12B	✓	50.4	48.6	−1.8	51.8	1.4	72.7	91.8	19.1	98.5	6.7
		×	36.5	34.2	−2.3	42.1	**5.6**	40.2	42.9	2.7	89.3	**46.4**

Table 2: Comparison of Direct, Chain-of-Thought, and Representational Probe accuracy across tasks, models, and subsets. R denotes the representation-probing accuracy from the **best-performing layer** of the language model. We report layer-wise performance in Appendix B. Δ columns show the gain of CoT over Direct (C−D) and of the representational probe over the stronger of the two baselines (R−max(D,C)). Bold entries indicate that the probing substantially exceeds verbal output for unknown entities.

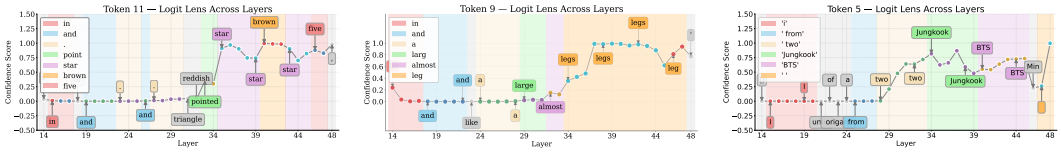
Beyond fine-grained correspondence: Art-Style Matching To test whether this effect extends to more holistic visual judgments, we additionally evaluate art-style matching (Fu et al., 2025), where models identify which target painting shares the reference painting's style. Art style depends on global properties such as texture, color usage, brushwork, and composition. We find that VLMs perform better on samples involving artists they can recognize. Recognizing the artist allows the model to avoid visually comparing the paintings' artistic styles. We discuss this experiment in Appendix I.

2.3 Logit Lens Reveals VLMs Explicitly Recover Semantic Anchors

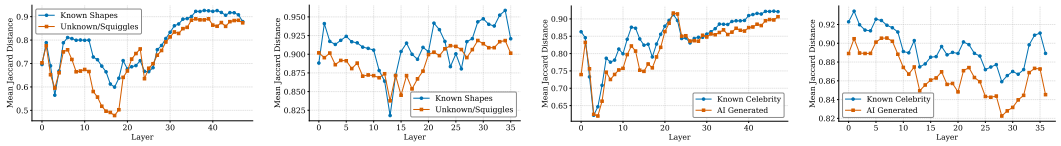
We use Logit Lens (nostalgebraist, 2020; 2021) to directly decode the hidden representations of visual tokens inside the language model, confirming that VLMs explicitly align known visual tokens to semantic labels. We refer to the top-1 decoded token as the *Logit Lens token* from here on. Since each entity is patched into multiple visual tokens, we approximate the semantic discernibility of two entities by measuring the difference between their Logit Lens token sets. Specifically, we compute the Jaccard Distance $D_J(A, B \mid L) = 1 - \frac{|\mathrm{LL}_L(\mathcal{V}_A) \cap \mathrm{LL}_L(\mathcal{V}_B)|}{|\mathrm{LL}_L(\mathcal{V}_A) \cup \mathrm{LL}_L(\mathcal{V}_B)|}$ where $\mathrm{LL}_L(\mathcal{V}_X)$ denotes the set of Logit Lens tokens for entity X's visual tokens at layer L.

Since each target image contains 4 entities, we compute the Jaccard Distance for all $\binom{4}{2}$ pairs per image and average over the entire dataset to obtain the *Mean Jaccard Distance* for each layer. A **higher** Jaccard Distance implies that the visual tokens of two entities are **more differentiable** in semantic space and therefore easier for VLMs to tell apart. We would expect known shapes and faces to surface more unique Logit Lens tokens and therefore higher Jaccard Distance. We note that Jaccard Distance **underestimates** semantic discernibility for three reasons: (1) we only use the top-1 decoded Logit Lens token; (2) it does not account for synonyms; and (3) since the shapes/faces are naturally similar, they share some Logit Lens tokens such as "polygon", "geometry", "eye", "nose", names of colors for shapes, and ethnicity for faces. Furthermore, prior work has shown that Logit Lens itself underestimates the interpretability of visual tokens in early layers (Krojer et al., 2026), since the unembedding matrix is trained to decode only the last layer. However, since our goal is to study the *relative* difference between known and unknown entities rather than absolute interpretability performance, Logit Lens suffices.

Results Figures 3a and 3c visualize the Logit Lens trajectory of individual visual tokens in Gemma3-12B for two Known Entities: a 2D star shape and a photograph of the celebrity. The decoded token evolves through distinct phases as it passes through the LM transformer



(a) Logit Lens for Known Shape (Star) (b) Logit Lens for Unknown Shape (c) Logit Lens for Known Face (Jungkook)



(d) Gemma - Shapes (e) Qwen - Shapes (f) Gemma - Faces (g) Qwen - Faces

Figure 3: Logit Lens analysis. **Top row:** layerwise decoded tokens for a known shape (star) and a known face (Jungkook) in Gemma3-12B, showing the progression from semantically unrelated tokens to exact labels and encyclopedic associations. For the unknown shape, the Logit Lens probe surfaces a hallucinated token ("legs"). **Bottom row:** Mean Jaccard Distance across layers for shapes and faces in Gemma3-12B and Qwen3VL-8B. Known entities (blue) yield consistently higher Mean Jaccard Distance than unknown entities (orange), confirming greater semantic discernibility in the hidden representations.

blocks. From Layer 0 to ∼30, the Logit Lens token is semantically unrelated to the visual input (e.g., "and," "in," "from") and carries near-zero probability mass. Around Layer ∼31–33, a transitional phase emerges, more pronounced for the shape, where approximate descriptors appear ("triangle," "reddish," "pointed" for the star in Fig. 3a). By Layer ∼34, the top-1 probability surges and the decoded token snaps to the correct semantic label: "star" and "Jungkook" in Figs. 3a and 3b, respectively. In the deepest layers, both tokens continue to evolve, accumulating higher-level associations: "five" for the star (reflecting the five-pointed shape) and "BTS" for the face (the celebrity's group). This layerwise progression, from noise, through approximate descriptors, to an exact semantic anchor, and finally to encyclopedic associations, demonstrates that VLMs explicitly recover discrete names for known visual entities within their intermediate representations. In contrast, Figure 3b shows that a visual token for an Unknown 2D shape spuriously maps to the "legs" token. We show additional examples in Appendix F.

Figures 3d–3g show that known shapes and faces have consistently higher Mean Jaccard Distance than unknown ones from Layer 4 onwards in both Gemma3-12B and Qwen3VL-8B. This confirms that Logit Lens produces more unique tokens for known entities and that the underlying hidden representations are more semantically discernible for the language model across both model families. Because semantically meaningful tokens emerge only after approximately Layer 30 in Gemma3-12B, the higher Mean Jaccard Distance observed in these later layers is particularly informative. Although the metric fluctuates considerably in earlier layers, it becomes consistently higher for known entities beyond Layer 30, indicating that their representations are more semantically distinguishable to the language model in the layers where meaningful visual semantics emerge.

3 Teaching Arbitrary Names

In Section 2, we showed that VLMs' performance on the semantic, shape, and face correspondence tasks is greatly affected when the model lacks a semantic label for the reference entity, despite the internal representations encoding enough visual information. In this section, we investigate whether teaching VLMs an arbitrary name for previously unknown shapes can close this gap on the shape correspondence task. Once a shape has a name, the VLM can use its default strategy of short-circuiting visual comparison, instead matching shapes by their learned labels: *"REF is John and C is John, so the answer is C."*

Experimental Setup We finetune Qwen3VL-2B and Gemma3-4B to teach them arbitrary names for each squiggle shape. We present the image containing a shape and ask, "What is the name of this object?" and train the model to generate the name. To prevent overfitting, we train the VLMs on a varied set of single-image tasks and aggressively augment the images. All tasks are single images with at most two shapes per image, and are deliberately designed to be distinct from the original 2D shape correspondence task, which had two images with four shapes per image. This ensures the VLM is not inadvertently trained to compare shapes between images. Examples of the tasks can be found in Appendix G.

We test three sets of names: ordinary object names that already map to a unique visual object (cup, anchor, feather, etc.), human names that are known to LLMs but lack a unique visual counterpart (John, Mary, Charles, etc.), and randomly generated 6-character strings that are completely unknown ("0QK2Z2", "5F1FT3", "OZ0W0M", etc.).

Model	Eval. Strategy	Name Set	Accuracy
Qwen3VL-2B	Rep. Probe		74.2
	VQA	Baseline	29.0
		Random	62.8
		Human	70.2
		Ordinary	86.0
Gemma3-4b	Rep. Probe		91.7
	VQA	Baseline	30.5
		Random	65.1
		Human	41.0
		Ordinary	50.5

Table 3: Direct VQA accuracy on unknown shape correspondence after learning arbitrary names. Rep. Probe (pre-finetuning) shown as reference.

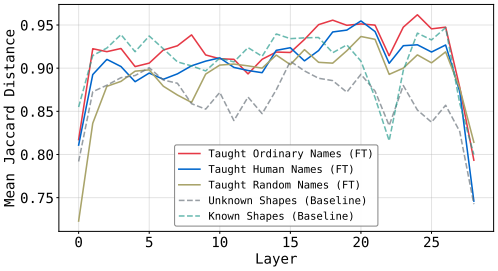


Figure 4: Mean Jaccard Distance across layers for Qwen3VL-2B after learning arbitrary names. Finetuned models close the gap between the unknown and the known shapes baselines.

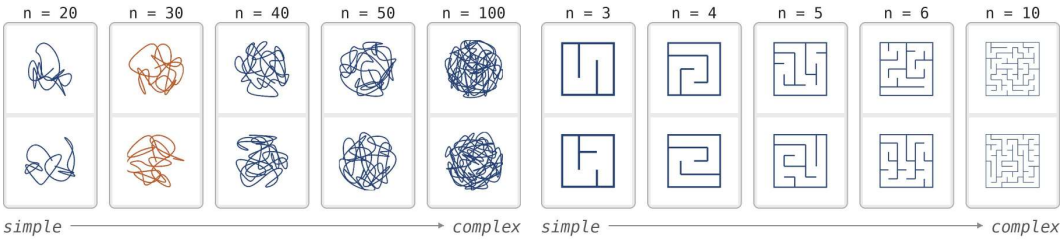
Results Table 3 shows the results of teaching VLMs arbitrary names for each unknown squiggle object. All three name types consistently improve VQA accuracy over the baseline.[2] For Qwen3VL-2B, ordinary names reach 86.0% (from 29.0% baseline), surpassing even the pre-finetuning Representation Probe of 74.2%. Human names reach 70.2% and random names 62.8%. Qwen3VL-2B benefits most from ordinary names (86.0%), while Gemma3-4B benefits most from random names (65.1%). The three name sets differ in average tokenization length (1 token for ordinary, ∼1.4 for human, ∼4.7 for random), which possibly interacts with how easily each model learns and uses them. Figure 4 confirms the mechanism: Logit Lens reveals that finetuned Qwen models close the Mean Jaccard Distance gap between unknown and known shape baselines, and the ordering (Ordinary > Human > Random) matches the VQA accuracy ordering, confirming a positive correlation between semantic discernibility and downstream performance. Inspecting the Chain-of-Thought reasoning confirms that the VLM **short-circuits** pixel comparison using the learned semantic anchors (see Appendix G.3). Instead of referring to fine-grained visual details, the VLM simply reasons *"REF is a brick and Choice D is a brick so the answer is D."*.

4 Is Semantic Alignment A Requirement or A Side Effect?

In this section, we investigate whether semantic anchoring is a requirement for good VLM performance or a side effect of VLM pretraining. Note that the need for semantic anchors would limit VLMs to the granularity of existing vocabulary, which is undesirable.

To test this, we finetune VLMs on shape correspondence using one set of shapes and evaluate on entirely different shapes the model has never seen, including a different shape family (mazes). A model that has learned to compare visual details directly should generalize across shape families. A model that relies on internally assigned semantic labels has no labels for novel shapes and should fail.

[2]We show the representation probing after finetuning in Appendix Figure 15.



(a) Generated Squiggles (b) Generated Mazes

Figure 5: Set of procedurally-generated shapes tested. n denotes the shape complexity variable: number of anchor points for squiggles and grid size for mazes. Orange denotes the train dataset (Squiggles n=30) and blue denotes the test dataset.

Experimental Setup We finetune Qwen3VL-2B and Gemma3-4B on shape correspondence using squiggles with medium complexity (n=30 anchor points). Our training dataset consists of 1000 synthetically generated image pairs of the type shown in Figure 2c and 2d. We evaluate on held-out squiggles with both simpler and more complex configurations, procedurally generated mazes whose rectilinear grid structure bears no geometric resemblance to the shapes seen during training, and on semantic and face correspondence tasks that represent extremely out-of-distribution domains.

		Squiggle			Maze		
Model	Domain Shift	n	Base Acc.	FT Acc.	n	Base Acc.	FT Acc.
Qwen3VL-2B	ID	30	29	99.3			
	OOD	20	36.7	100	3	36.8	99.0
		30	28.8	98.7	4	34.2	99.4
		40	27.8	93.0	5	28.6	90.5
		50	25.9	76.1	6	25.9	82.2
		100	28.7	61.4	10	24.0	62.8
Gemma3-4b	ID	30	30.3	100			
	OOD	20	36.1	100	3	36.3	99.9
		30	31.9	98.7	4	37.3	99.1
		40	26.7	94.0	5	31.9	99.3
		50	27.2	82.7	6	28.6	97.1
		100	27.0	63.2	10	28.0	81.1

Table 4: Direct VQA accuracy on shape correspondence after finetuning on squiggles (n=30). Base Acc. is pre-finetuning. FT Acc. is after finetuning.

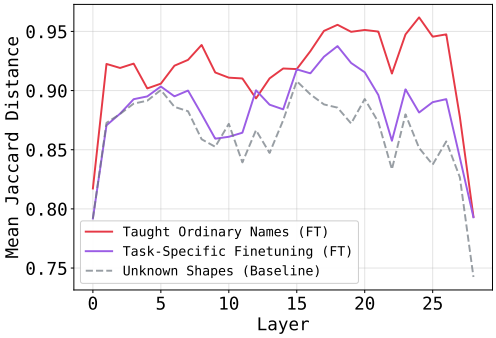


Figure 6: Task-specific finetuning has lower Jaccard Distance than teaching ordinary names, despite higher VQA accuracy (98.7% vs 86.0%).

Results Table 4 shows that both Qwen3VL-2B and Gemma3-4B exhibit strong OOD generalization after finetuning, achieving near-perfect accuracy on held-out squiggles of similar or lower complexity, with graceful degradation as complexity increases. Critically, they also generalize to mazes despite being trained only on squiggles, with Gemma3-4B reaching 99.3% on complex $n = 5$ maze structures. This rules out the internal labeling hypothesis: VLMs generalize to unseen shapes, including an entirely different shape family, indicating they have learned visual comparison rather than memorizing specific shapes.

We further find that finetuning on squiggles improves performance on face and semantic correspondence tasks, too, which are extremely out-of-distribution with respect to squiggles. Finetuned Qwen3VL-2B improves by 16% on face correspondence and 10.7% on semantic correspondence. Finetuned Gemma3-4B improves by 8.7% and 5.8% on face and semantic correspondence, respectively. The full results can be found in Appendix H.

Figure 6 reveals that task-specific finetuning has a lower Mean Jaccard Distance than teaching ordinary names (Section 3), but it achieves higher VQA accuracy (98.7% vs 86.0%). This suggests that task-specific finetuning teaches the VLM to leverage fine-grained visual information through a mechanism beyond semantic discernibility. Prior work has shown that task-specific finetuning improves VLM performance on visual correspondence tasks (Fu et al., 2025; Liu et al., 2025), but the underlying mechanism was unclear. Our results suggest

semantic anchoring is a sufficient but not necessary condition: VLMs can also learn to use visual information directly when the training task demands fine-grained visual comparison, and the two pathways (semantic anchoring in Section 3 vs. direct visual comparison here) represent distinct mechanisms for closing the "hidden-in-plain-sight" gap.

5 Related Work

Recent work has shown that current VLMs still struggle on tasks that require fine-grained visual reasoning. These limitations appear in chart and diagram understanding, optical illusions, and subtle visual discrimination, where success depends on preserving local evidence and comparing small visual differences rather than relying on coarse object semantics (Zhang et al., 2024; Shahgir et al., 2024; Chen et al., 2026; Kim et al., 2026; Gu et al., 2025; Guo et al., 2025; Nath et al., 2025; Danish et al., 2025). A closely related line of work further shows that the problem is often not the absence of visual information itself. On perception-heavy tasks, the needed signal can still be recovered from internal representations even when the model's final text answer is wrong (Fu et al., 2025; Liu et al., 2025). Together, these findings suggest that the main bottleneck is not only visual encoding, but also the model's limited ability to preserve and use fine-grained visual evidence during decoding.

Saini et al. (2026) show that for the attention mechanism of VLMs to yield mathematically meaningful similarity scores between visual queries and textual keys (or vice versa), both modalities must operate within a shared, dimensionally consistent coordinate system. Indeed, several studies suggest that visual features are progressively transformed into more language-like states inside the language model. Across layers, image-conditioned hidden states become more interpretable in vocabulary space, some query tokens capture global image information, and fine-grained attributes can still be recovered from spatially localized visual tokens (Neo et al., 2024; Kaduri et al., 2025; Wang et al., 2025). Layer-wise analyses of intermediate representations make it possible to track this transformation more directly and relate internal visual states to later textual predictions (Neo et al., 2024; Kaduri et al., 2025; Liu et al., 2025). Other work shows that the language decoder can compensate for weak or incomplete visual representations, suggesting a division of labor between visual features and language priors (Takishita et al., 2025; Merullo et al., 2022). The same mechanism also affects downstream behavior: visual inputs can be harder than textual inputs for factual recall, and even when visual and textual tasks are matched, they can rely on partly different internal circuits (Cohen et al., 2025; Nikankin et al., 2025).

Our work connects these two lines of research. We argue that one reason pixel-level tasks remain difficult is that VLMs perform best when the target can be mapped to a clear semantic anchor in language space. We study this in correspondence-style tasks, where success depends on matching local regions or shapes rather than naming an object category. Across semantic, shape, and face correspondence, we show that the representation-output gap becomes much larger for unnameable entities even when internal features remain sufficient. We further compare two ways of closing this gap: teaching arbitrary names, which strengthens a language-mediated shortcut, and direct task finetuning, which instead improves visual comparison without requiring the same kind of semantic recoding.

6 Conclusion

We propose a simple explanation for the hidden-in-plain-sight gap: VLMs are much better at using visual information when they can attach a semantic label to it. Across semantic, synthetic, and face correspondence tasks, performance is substantially higher for familiar, nameable entities than for semantically unknown ones, even when representations retain the information needed to solve the task. Logit Lens analyses and naming interventions suggest that semantic labels let models shortcut direct visual comparison by mapping the problem to language. At the same time, semantic labeling is not necessary for strong performance. With direct task finetuning, VLMs can learn more general visual skills that transfer to unseen shapes and families while relying less on semantic anchoring. Together, our results suggest that many current VLM failures reflect a learned shortcut rather than a fundamental limitation of multimodal reasoning.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Y. charles, Yiping Bao, Yuantao Fan, Guopeng Li, Haiyang Shen, Xuanzhong Chen, Wendong Xu, Shuzheng Si, Zefan Cai, Wenhao Chai, Ziqi Huang, Fangfu Liu, Tianyu Liu, Baobao Chang, Xiaobo Hu, Kaiyuan Chen, Yixin Ren, Yang Liu, Yuan Gong, and Kuan Li. Babyvision: Visual reasoning beyond language, 2026. URL https://arxiv.org/abs/2601.06521.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems, 2026. URL https://arxiv.org/abs/2505.11831.

Ido Cohen, Daniela Gottesman, Mor Geva, and Raja Giryes. Performance gap in entity knowledge extraction across modalities in vision language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29095–29108, 2025.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. GEOBench-VLM: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111:98 – 136, 2014. URL https://api.semanticscholar.org/CorpusID:207252270.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations, 2025. URL https://arxiv.org/abs/2506.08008.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015. URL https://arxiv.org/abs/1508.06576.

Google DeepMind. Nano banana 2: Google's image generation and editing model, 2025. URL https://deepmind.google/models/gemini-image/. Accessed: 2026-03-13.

Difei Gu, Yunhe Gao, Mu Zhou, and Dimitris Metaxas. Anatomy-VLM: A fine-grained vision-language model for medical interpretation. *arXiv preprint arXiv:2511.08402*, 2025.

Zhenhao Guo, Rachit Saluja, Tianyuan Yao, Quan Liu, Yuankai Huo, Benjamin Liechty, David J. Pisapia, Kenji Ikemura, Mert R. Sabuncu, Yihe Yang, and Ruining Deng. Glo-VLMs: Leveraging vision-language models for fine-grained diseased glomerulus classification. *arXiv preprint arXiv:2508.15960*, 2025.

Raul Ismayilov, Dzemila Sero, and Luuk Spreeuwers. Fluxsynid: A framework for identity-controlled synthetic face generation with document and live images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3698–3708, 2025.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14549–14558, 2025.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert, 2020. URL https://arxiv.org/abs/2004.12832.

Minkyu Kim, Sangheon Lee, and Dongmin Park. VLM-SubtleBench: How far are VLMs from human-level subtle comparative reasoning? *arXiv preprint arXiv:2603.07888*, 2026.

Benno Krojer, Shravan Nayak, Oscar Mañas, Vaibhav Adlakha, Desmond Elliott, Siva Reddy, and Marius Mosbach. Latentlens: Revealing highly interpretable visual tokens in llms. *arXiv preprint arXiv:2602.00462*, 2026.

Benlin Liu, Amita Kamath, Madeleine Grunde-McLaughlin, Winson Han, and Ranjay Krishna. Visual representations inside the language model. *arXiv preprint arXiv:2510.04819*, 2025.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019.

Vishwesh Nath et al. VILA-M3: Enhancing vision-language models with medical expert knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. *arXiv preprint arXiv:2506.09047*, 2025.

nostalgebraist. Interpreting GPT: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. LessWrong.

nostalgebraist. Logit lens on non-GPT2 models + extensions. https://colab.research.google.com/drive/1MjdfK2srcerLrAJDRaJQKO0sUiZ-hQtA, 2021. Google Colab notebook.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4624–4633, 2019.

Harshvardhan Saini, Samyak Jha, Yiming Tang, and Dianbo Liu. When language overwrites vision: Over-alignment and geometric debiasing in vision-language models, 2026. URL https://arxiv.org/abs/2605.08245.

Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. Illusionvqa: A challenging optical illusion dataset for vision language models. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2403.15952.

Sho Takishita, Jay Gala, Abdelrahman Mohamed, Kentaro Inui, and Yova Kementchedjhieva. LLMs can compensate for deficiencies in visual representations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15253–15272, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.825. URL https://aclanthology.org/2025.findings-emnlp.825/.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2505.23941.

Sudong Wang, Yunjian Zhang, Yao Zhu, Jianing Li, Zizhe Wang, Yanwei Liu, and Xiangyang Ji. Towards understanding how knowledge evolves in large vision-language models, 2025. URL https://arxiv.org/abs/2504.02862.

Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. *IEEE Winter Conference on Applications of Computer Vision*, pp. 75–82, 2014. URL https://api.semanticscholar.org/CorpusID:11266650.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024. URL https://arxiv.org/abs/2403.14624.

Supplementary Material: Appendices

A Qualitative Chain-of-Thought Examples on Named and No-Name Semantic Correspondence

Qwen3VL 2B's CoT Response on **<<Name>>** Subset of Semantic Correspondence Task

Qwen3VL-2B's CoT:

To determine the correct correspondence, we need to analyze the reference point and find the point on the second image that has the same semantic meaning. The reference point 'REF' in the first image is located on the **front of the aircraft's nose**, specifically on the **nose cone**. It is positioned at the very front of the aircraft's fuselage, just above the cockpit windows.

- Now, let's examine the points on the second image:
  - **Point A** is on the **front landing gear** (the wheel assembly) of the aircraft.
  - **Point B** is on the **top of the aircraft's fuselage**, near the cockpit.
  - **Point C** is on the **front of the aircraft's nose cone**.
  - **Point D** is on the **top of the cockpit windows**.

The key is to identify the point that shares the same semantic role as the reference point. The reference point is on the nose of the aircraft. The point that is also on the nose of the aircraft is **Point C**. Therefore, the correct answer is (C) Point C.

Qwen3VL 2B's CoT Response on **<<No-Name>>** Subset of Semantic Correspondence Task

Qwen3VL-2B's CoT:

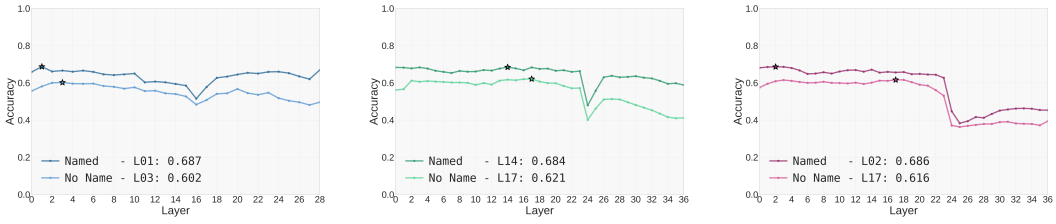
To solve this problem, we need to identify the corresponding point on the second image based on the reference point in the first image. The reference point 'REF' in the first image is located on the **top of the rear window** of the silver car.

- In the second image, we have four points on the blue car:
  - **A** is on the **rear wheel**.
  - **B** is on the **top of the rear window**.
  - **C** is on the **rear taillight**.
  - **D** is on the **top of the rear window**.

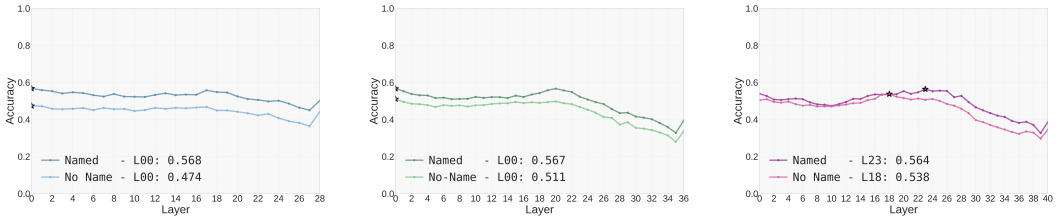
The reference point 'REF' is on the top of the rear window. Both 'B' and 'D' are on the top of the rear window. Since the reference point is on the top of the rear window, the corresponding point on the second image must also be on the top of the rear window. Therefore, the correct choice is (B) Point B.

Figure 7: Chain-of-Thought reasoning allows the VLM to recover the semantic label for each point explicitly and essentially convert the task to a verbal problem. However, when the points lack a concrete semantic label, the transcription process becomes harder, leading to hallucinations, as shown by the VLM mislabeling point B as the top of the rear window instead of the bottom.

B Representation Probing Performance Per Layer For All Tasks

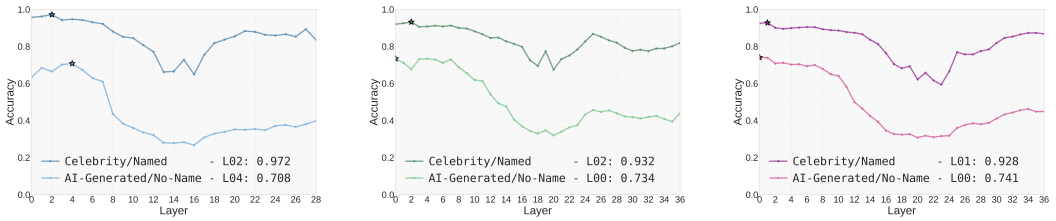


(a) Qwen3VL-2B (b) Qwen3VL-4B (c) Qwen3VL-8B

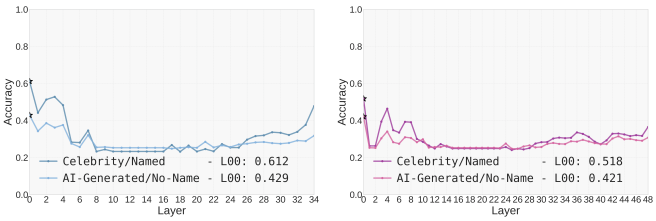


(d) InternVL3.5-2B (e) InternVL3.5-8B (f) InternVL3.5-14B

Figure 8: Layer-wise Representation Probing Performance for Nameable and No-Name reference points from the **Semantic Correspondence** (Min et al., 2019) task.

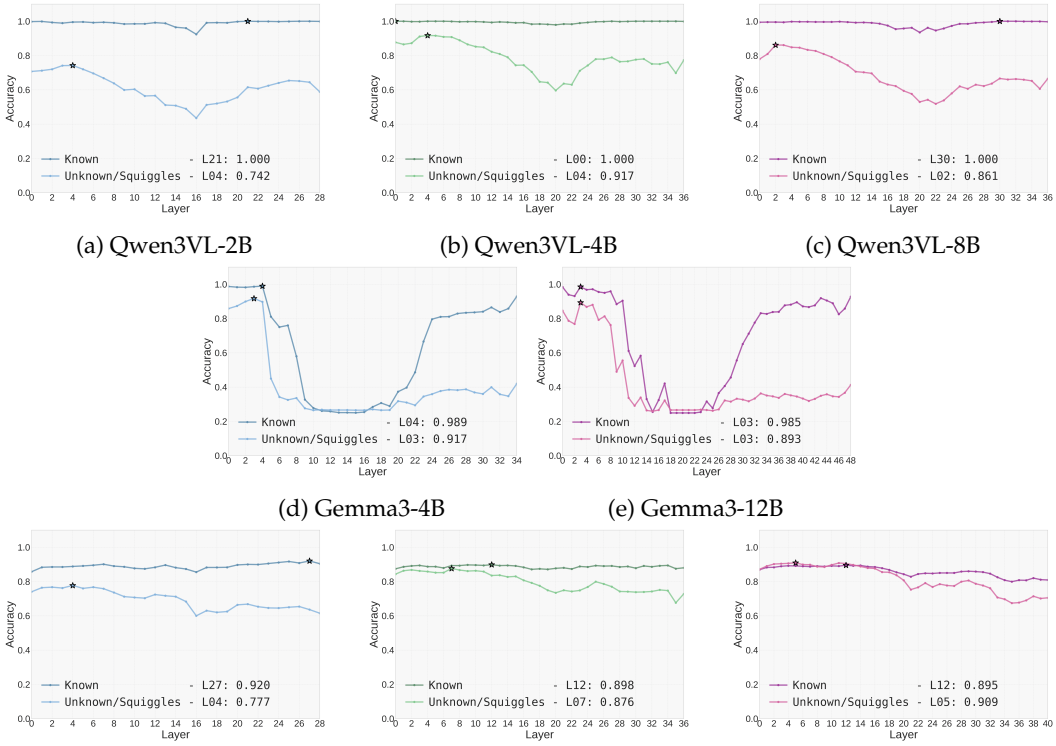


(a) Qwen3VL-2B (b) Qwen3VL-4B (c) Qwen3VL-8B



(d) Gemma3-4B (e) Gemma3-12B

Figure 9: Layer-wise Representation Probing Performance for **Known (Celebrity) and Unknown (AI-Generated) Faces**.



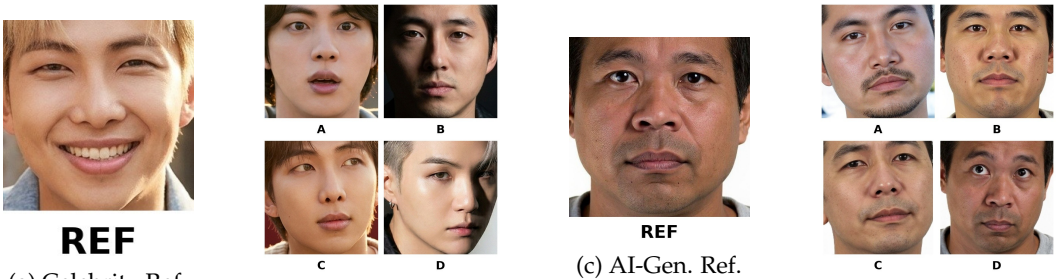
(a) Qwen3VL-2B (b) Qwen3VL-4B (c) Qwen3VL-8B

(d) Gemma3-4B (e) Gemma3-12B

(f) InternVL-3.5-2B (g) InternVL-3.5-8B (h) InternVL-3.5-14B

Figure 10: Layer-wise Representation Probing Performance for **2D Known and Procedurally Generated (Squiggle) Shapes**.

C Faces Dataset Experimental Setup



(a) Celebrity Ref. (b) Celebrity Tgt. (c) AI-Gen. Ref. (d) AI-Gen. Tgt.

Figure 11: Face correspondence task. Known celebrities (left) and AI-generated faces (right).

To build our face correspondence dataset, we collect an initial set of celebrities and use Nano-Banana-2 (Google DeepMind, 2025) to generate four images per person, ensuring uniform resolution across all samples and avoiding artifacts that may arise from real images scraped from the internet. We construct model-specific splits by probing each VLM's knowledge: given an image, we prompt the model with "Who is the person in this image?" and designate a person as known to the model if at least one of the four generated images is identified correctly. Table 5 shows the initial set and model-specific subsets of celebrities.

VLM	Recognized Faces
All	Bruce Lee, BTS J-Hope, BTS Jimin, BTS Jin, BTS Jungkook, BTS RM, BTS Suga, BTS V, Jack Ma, Jackie Chan, Masayoshi Son, Song Heungmin, Steven Yeun, Yao Ming
Qwen-2B	BTS Jimin, BTS Jin, BTS Jungkook, BTS Suga, BTS V, Bruce Lee, Jack Ma, Jackie Chan, Song Heungmin, Yao Ming
Qwen-4B	BTS J-Hope, BTS Jimin, BTS Jin, BTS RM, BTS Suga, BTS V, Bruce Lee, Jack Ma, Jackie Chan, Steven Yeun, Yao Ming
Qwen-8B	BTS Jimin, BTS Jin, BTS Jungkook, BTS Suga, BTS V, Bruce Lee, Jack Ma, Jackie Chan, Steven Yeun, Yao Ming
Gemma-4B	BTS Jimin, BTS Jin, BTS Jungkook, BTS RM, BTS V, Bruce Lee, Jack Ma, Jackie Chan, Masayoshi Son, Steven Yeun
Gemma-12B	BTS Jimin, BTS Jin, BTS Jungkook, BTS RM, BTS Suga, BTS V, Jackie Chan, Steven Yeun

Table 5: VLM-specific Face Correspondence dataset.

Nano Banana 2 Prompt `{Reference Image} Generate four images of this person with varying expressions, lighting, and poses. Also, change his clothes and hairstyles. Ensure the images are square, close-ups of his face. The person should not look too far to one side. The face should be almost fully visible. There should be no visible hands in the image.`

D InternVL3.5's Facial Recognition Failure and Results on Shape Correspondence

Facial Recognition Failures Our pilot experiments revealed that the InternVL3.5 model family fails to recognize the images of most celebrities. We probe the VLM with the prompt *"Who is the person in this image? Respond with the format: 'The person in the image is ¡name¿.' with no other text."*. Table 6 shows that InternVL-3.5-14B fails to recognize all but two celebrities. InternVL-3.5-2B and 8B are similarly limited. As such, we did not evaluate InternVl-3.5 on the facial correspondence task.

Ground Truth	InternVl3.5 14B Predictions (Images 1–4)	Correct
Barack Obama	Barack Obama, Barack Obama, Barack Obama, Barack Obama	4/4
Cristiano Ronaldo	*IDK*, Cristiano Ronaldo, Paul Walker, *IDK*	1/4
Donald Trump	Donald Trump, Donald Trump, Donald Trump, Donald Trump	4/4
Lionel Messi	Cristiano Ronaldo, Cristiano Ronaldo, Chris Hemsworth, Aaron Hernandez	0/4
BTS J-Hope	Park Yohan, Park Yoo-chun, Jungkook, Lee Min-ho	0/4
BTS Jimin	Lee Seung-gi, Yohan Lemoine, Park Yohan, Park Yohan	0/4
BTS Jin	Park Yoo-chun, Park Bo-gum, Park Yohan, Park Yoo-chun	0/4
BTS Jungkook	Lee Seung-gi, Park Yoo-chun, Park Yohan, Park Yoo-chun	0/4
BTS RM	Park Yohan, Lee Hong-gi, Lee Jong-suk, Lee Min-ho	0/4
BTS Suga	Park Yohan, Park Yoo-chun, Park Yohan, Lee Jong-suk	0/4
BTS V	Park Yohan, Lee Seung-gi, Park Yohan, Lee Min-ho	0/4
Bruce Lee	Hu Ge, Tony Leung, Lee Min-ho, Tony Leung	0/4
Jack Ma	*IDK*, *IDK*, Chen Songlian, *IDK*	0/4
Jackie Chan	*IDK*, *IDK*, *IDK*, Chen Kaige	0/4
Jet Li	Tony Leung, *IDK*, Tony Leung, *IDK*	0/4
Masayoshi Son	Chen Songlian, Masahiro Hamaguchi, Tetsuro Yamada, Chen Songlian	0/4
Shigeru Miyamoto	Tetsuro Ishida, Masahiro Miki, Chen Kaige, Tetsu Komai	0/4
Shohei Ohtani	Aaron Yonda, Lee Jong-suk, Aaron Kwok, Aaron Yonda	0/4
Son Heung-min	Lee Min-ho, Jun, Lee Jong-suk, Chen Kun	0/4
Steven Yeun	Aaron Kwok, Chen Liang, Lee Byung-hun, Shuichi Saihara	0/4
Yao Ming	Xiaoyu Wang, Aaron Kwok, Joonas Suotamo, Wang Zhi	0/4

Table 6: InternVL 3.5 14B face recognition probing results. *IDK* means the VLM responded that it was unable to identify the person. We test additional faces of very famous celebrities to confirm InternVL-3.5 failure in face recognition.

InternVL-3.5 on Shape Correspondence We evaluated the InternVL-3.5 model family on shape correspondence since they recognize common geometric shapes ("square", "circle", "star", etc). Table 7 shows that InternVL-3.5 has a larger gap between representation probing and verbal output for unknown shapes. Furthermore, Chain-of-Thought reasoning is only

beneficial when the shapes are known, completely corroborating our findings for Gemma3 and Qwen3VL in Table 2.

Model		Known	2D Shapes				
			Direct (D)	CoT (C)	C−D Δ	Rep. Probe (R)	R−max(D,C) Δ
InternVL 3.5	2B	✓	46.1	76	29.9	92	16
		×	26.1	27.7	1.6	77.7	**50**
	8B	✓	66.4	89.9	23.5	89.8	-0.1
		×	27.7	27.6	-0.1	86.8	**59.1**
	14B	✓	55.1	90.9	35.8	89.5	-1.4
		×	32.6	31.4	-1.2	90.4	**57.8**

Table 7: Comparison of Direct, Chain-of-Thought, and Representational Probe accuracy for InternVL-3.5. R denotes the representation-probing accuracy from the **best-performing layer** of the language model. We report layer-wise performance in Appendix Fig. 10. Δ columns show the gain of CoT over Direct (C−D) and of the representational probe over the stronger of the two baselines (R−max(D,C)). Bold entries indicate that the probe substantially exceeds the best prompted baseline for Unknown 2D-shapes.

E Gemma-3 Performance on Semantic Correspondence

Model	Size	Subset	Direct (D)	CoT (C)	C−D Δ	Rep. Probe (R)	R−max(D,C) Δ
Gemma-3	4B	Named	25.7	26.3	0.6	38.6	**12.3**
		No-Name	24.8	26.2	**1.4**	34.6	8.4
	12B	Named	31.2	35.9	**4.7**	39.4	3.5
		No-Name	27.5	28.4	0.9	33.5	**5.1**

Table 8: Semantic correspondence results for Gemma-3 on SPairs71k. Gemma-3-4B performs near chance and therefore does not exhibit the trends observed in more capable VLMs. In contrast, Gemma-3-12B shows a larger gap between Representation Probing and the best textual strategy, **R−max(D, C)**, on No-Name keypoints, consistent with the pattern observed for the Qwen3-VL and InternVL3.5 families in Table 1.

F Additional Logit Lens Results

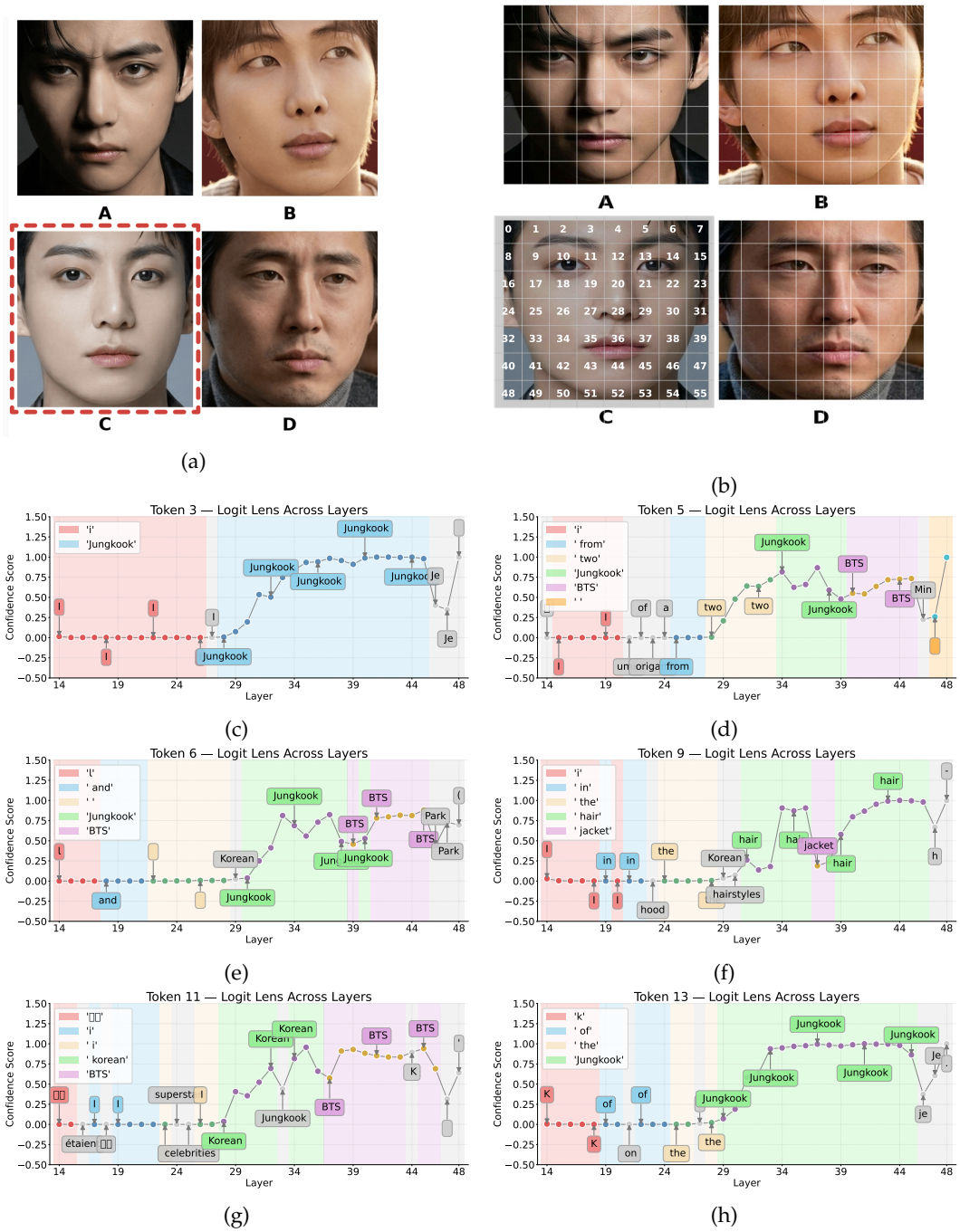


Figure 12: Qualitative Logit Lens example from Gemma-3-12B. The results show that some decoded tokens are highly informative of the person's identity and affiliations, while Fig. 12f shows an example of a visual token that encodes generic semantic information.

Figure 13: All Gemma3-12B Logit Lens tokens for (Unknown Shape 1).

Figure 14: All Gemma3-12B Logit Lens tokens for . (Unknown Shape 2)

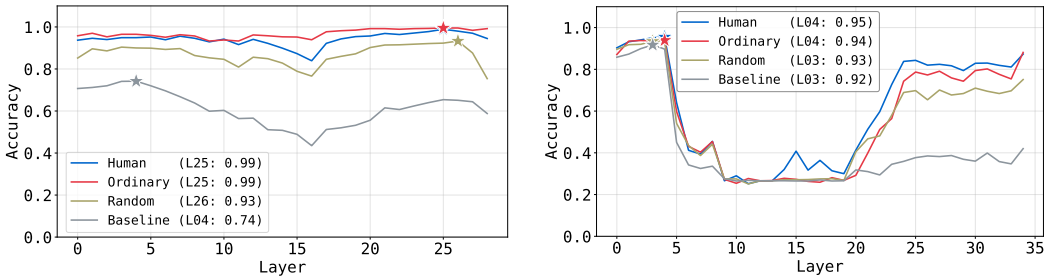
G Teaching Arbitrary Names To VLMs

Name Set	Names
Random	0QK2Z2, 5F1FT3, OZ0W0M, ALCTDF, DNXXB0, ION17F, K0XQNF, UTNWY7, JT1GWQ, 1VZS0M
Human	John, Mary, Charles, Elizabeth, William, Margaret, James, Catherine, Robert, Dorothy
Ordinary	cup, brick, anchor, fork, bell, shield, blade, horn, nest, arrow

G.1 Finetuning Tasks for Teaching Names

Task	Image Type	Example Prompt	Expected Answer
Naming	Single Image	"What is this shape called?"	Mary
Yes/No	Single Image	"Is this a cup?"	Yes.
Choice	Single Image	"Which of the following: fork, cup, bell, arrow?"	cup
Comparison	Single Image (Two Shapes)	"Which object is the arrow, A or B?"	A
Description	Single Image	"Can you describe the shape called John?"	A name derived from the Latin 'Iohannes'...

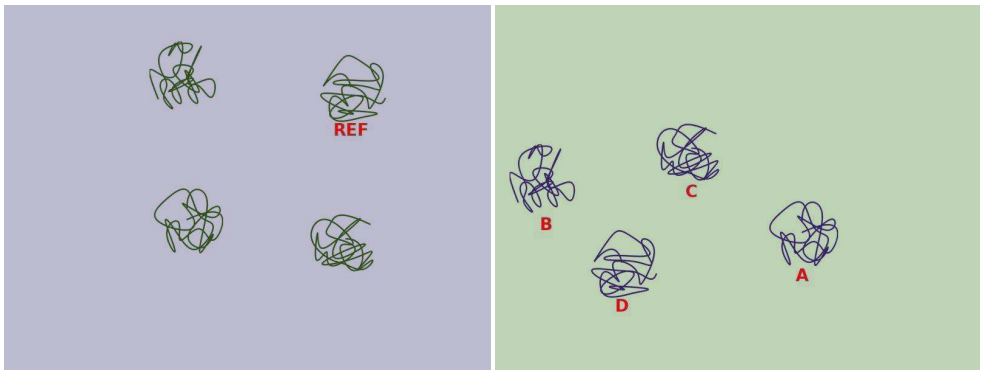
G.2 Representation Probing After Teaching Names



(a) Qwen3-VL 2B (b) Gemma3-4B

Figure 15: Representation Probing accuracy increases after teaching VLM names through finetuning. For Qwen, the improvement is consistent across all layers, matching the gap reported in Figure 10a. For Gemma, there is a sharp decline for all models after Layer 4, followed by a sharp recovery at Layer 19. This matches the behavior for Known shapes shown in Figure 10d.

G.3 Chain-of-Thought Example of Name-Trained Qwen3VL



Baseline Qwen3VL 2B's CoT Response:

To solve this, we need to find the scribble in the second image that has the same structural shape as the scribble marked "REF" in the first image. Let's analyze:

- The "REF" scribble in the first image is a complex, tangled mass of lines, with a distinct shape that includes a prominent upward-pointing "roof-like" structure at the top, and a more compact, dense cluster below it.

Now, let's examine the options in the second image:
- **A**: This scribble is a dense, tangled mass, but it lacks the distinct upward-pointing "roof-like" structure. Its shape is more circular and compact.
- **B**: This scribble is very similar to the "REF" scribble. It has a complex, tangled structure, and most importantly, it has a clear upward-pointing "roof-like" structure at the top, just like "REF". The overall shape and density of the lines match.
- **C**: This scribble is a simpler, less dense tangled mass. It lacks the prominent "roof-like" structure and is more open.
- **D**: This scribble is also a dense, tangled mass, but it doesn't have the distinct upward-pointing "roof-like" structure. It's more like a blob without a defined top shape.

The key is the structural similarity, not the position. The "REF" scribble has a specific shape: a tangled mass with a "roof-like" top. Only **B** matches this exact structural pattern. Therefore, the correct answer is **(B) Point B.**

Qwen3VL 2B's CoT Response After Being Taught Ordinary Object Names:

Let's break down the problem step by step.
1. **Identify the reference point in the first image.**
   - The reference point 'REF' is located on the top-right object in the first image. This object is a stack of bricks, which is a rectangular block with a roof.
2. **Analyze the objects in the second image.**
   - We have four choices: A, B, C, and D.
   - Choice A is a arrow.
   - Choice B is a shield.
   - Choice C is a anchor.
   - Choice D is a stack of bricks.
3. **Match the reference object to the correct choice.**
   - The reference object in the first image is a stack of bricks.
   - The object in the second image that matches this description is choice D.

Therefore, the correct answer is D.

Figure 16: Chain-of-thought example on unknown shapes. Without a semantic anchor, the VLM generates hallucinated descriptions that actively mislead its final answer.

H Extreme OOD Test After Task-Specific Finetuning With Squiggles

	Faces			Sem. Corr.		
Model	Split	Base Acc.	FT Acc.	Split	Base Acc	FT Acc
Qwen3VL-2B	Known	77.1	89.8	Named	36.6	52.3
	Unknown	41.1	60.4	No-Name	32.4	38.1
Gemma3-4b	Known	49.8	60.8	Named	26.1	35.50
	Unknown	32.4	38.8	No-Name	25.7	27.80

Table 9: Semantic and face correspondence performance after finetuning on squiggles (n=30). Base Acc. is the pre-finetuning baseline. FT Acc. is after finetuning.

H.1 Transfer to Open-Ended Visual Change Description

To test whether correspondence finetuning transfers beyond the multiple-choice format used during training, we evaluate Qwen3VL-2B on an open-ended visual change-description task derived from CLEVR-Change (Park et al., 2019). CLEVR-Change contains pairs of synthetic scenes with controlled changes in object attributes and spatial configurations, making it visually distinct from the procedurally generated squiggles used for finetuning. We prompt the model: *"Inspect the shapes, their colors, their positions, and their sizes in the two images. There are subtle differences between the two images. What are they?"* Because this is an open-ended generation task, we use Qwen3-8B as an automatic judge. The judge receives the model response together with the set of valid ground-truth difference descriptions and assigns a score from 0 to 10 based on how well the response matches the ground truth.

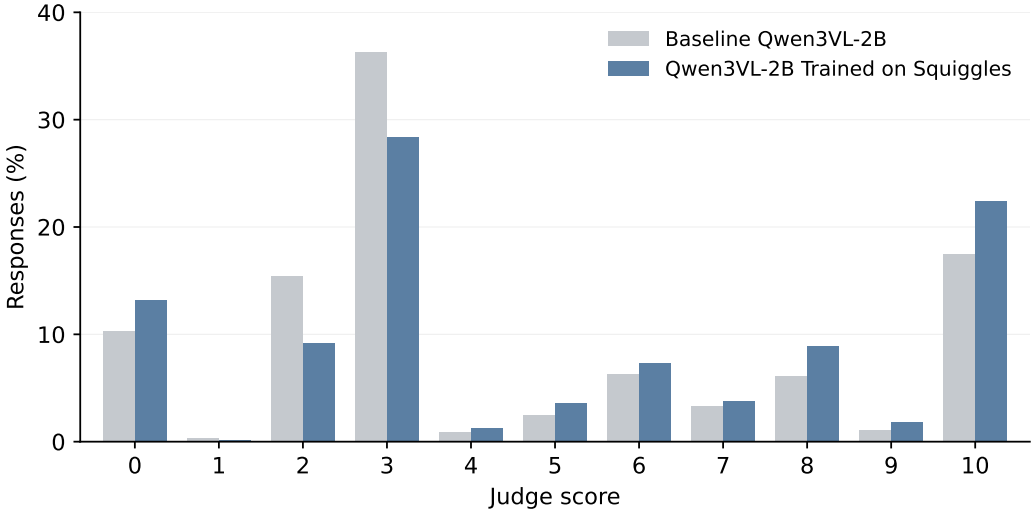


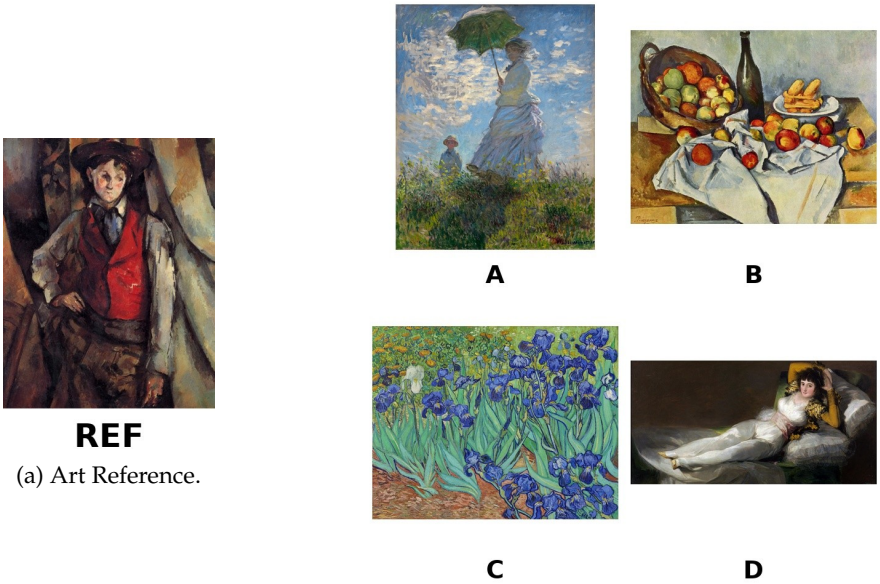
Figure 17: Distribution of Qwen3-8B judge scores on open-ended CLEVR-Change responses produced by Qwen3VL-2B before and after finetuning on squiggle correspondence. The finetuned model achieves a higher mean judge score (5.087 vs. 4.507) and produces more high-scoring responses (scores 8–10: 33.1% vs. 24.7%).

Figure 17 shows that finetuning shifts the score distribution toward higher-quality responses. The mean judge score improves from 4.507 to 5.087, and the fraction of high-scoring responses (scores 8–10) increases from 24.7% to 33.1%.

This experiment differs from the training task in both visual domain and output format. The model is not selecting among annotated candidates or producing a fixed label, but instead must generate a free-form description of a subtle change. The improvement therefore

provides evidence that correspondence finetuning does not merely teach the multiple-choice answer format, but transfers to a broader visual comparison behavior.

I Art Style Matching



(a) Art Reference.

(b) Art Style Option Panel.

Figure 18: Art Style task example. The ground truth answer is B — both the reference portrait and option B are by Paul Cézanne.

Our main experiments study semantic anchoring in tasks that require matching specific objects, object parts, shapes, or identities. We additionally test whether the same phenomenon appears in a more holistic visual task: *art style matching*. Unlike fine-grained correspondence, this task requires comparing global properties such as texture, color usage, brushwork, and composition. However, when the painter is recognizable, a VLM may bypass this visual comparison by identifying the painter of each artwork and matching the resulting names.

Dataset construction. We collect artworks from painters with visually distinctive styles, including Goya, Rembrandt, and Monet. For each VLM, we first probe whether it recognizes the painter of each artwork using the prompt: *"Which painter created this painting?"* We use the resulting predictions to construct model-specific *Known* and *Unknown* subsets. In the Known subset, the model can identify the relevant painters, whereas in the Unknown subset it cannot.

Each evaluation sample contains one reference artwork and a target image containing four candidate artworks. We prompt the model: *"Which painting in the second image has the same art style as the painting in the first image?"* We treat two artworks as having the same style when they were created by the same painter. This is an approximation, since an artist's style may vary across works and periods. We mitigate this limitation by selecting painters with visually distinctive styles and using multiple artworks with varied subject matter.

Evaluation. We evaluate textual performance using the same Direct and Chain-of-Thought (CoT) setups. We also evaluate the information available in the models' internal representations using two training-free comparison strategies. First, we apply the MaxSim representation probe used in our correspondence experiments. Second, we compare the representations using Gram-matrix distance, which captures correlations between features and has com-

Model	Subset	Textual Evaluation		Representation Probing	
		Direct	CoT	MaxSim	Gram Dist.
Qwen3VL-2B	Known	**65.4**	**59.4**	**65.2**	**62.0**
	Unknown	48.6	45.8	51.4	51.6
Qwen3VL-8B	Known	**72.0**	**73.0**	**68.2**	**62.6**
	Unknown	47.8	37.2	51.2	51.8
Gemma3-4B	Known	35.6	37.8	**53.6**	**47.8**
	Unknown	**36.0**	**42.4**	40.0	33.8
Gemma3-12B	Known	**58.0**	**56.0**	**55.0**	**55.2**
	Unknown	39.2	40.6	31.8	32.2

Table 10: Art style matching accuracy on model-specific Known and Unknown subsets. Known artworks are created by painters that the corresponding VLM can identify, while Unknown artworks are created by painters that it does not identify. Bold indicates the better-performing subset for each model and evaluation strategy.

monly been used to represent visual style (Gatys et al., 2015). For each representation-based method, we report the accuracy from the best-performing language-model layer.

Results. Table 10 shows that the effect of semantic anchors extends beyond matching individual objects and local visual details. Qwen3VL-2B, Qwen3VL-8B, and Gemma3-12B perform substantially better on Known artworks than on Unknown artworks under both textual evaluation strategies. The largest textual gap occurs for Qwen3VL-8B, whose CoT accuracy decreases from 73.0% on Known artworks to 37.2% on Unknown artworks. This is consistent with the model identifying recognizable painters and converting the visual style-matching problem into a comparison between semantic labels.

The representation-based evaluations generally show the same pattern. For Qwen3VL-2B, Qwen3VL-8B, and Gemma3-12B, both MaxSim and Gram-matrix distance perform better on known artworks. This suggests that recognizable artworks also become more distinguishable within the models' internal representations. Gemma3-4B is the main exception: its textual performance is weak and does not show a Known-over-Unknown advantage, although both representation-based strategies still perform substantially better on the Known subset. As in our other experiments, smaller models may lack the baseline capability required to reliably exploit semantic anchors through textual generation.

Unlike the correspondence experiments, representation probing does not consistently outperform the textual strategies on art style matching. For example, Qwen3VL-8B obtains 73.0% with CoT but only 68.2% with MaxSim. This does not necessarily indicate that the relevant information is absent from the hidden representations. Rather, MaxSim and Gram-matrix distance are fixed, training-free comparison functions and may not capture the holistic and spatially distributed features required to distinguish artistic styles. A learned probe could potentially extract this information more effectively.

Overall, these results provide additional evidence that semantic anchoring is not limited to fine-grained correspondence. Even for a task requiring holistic visual comparison, VLM performance is generally stronger when the visual entities can first be mapped to recognizable language concepts.