# Anchored, Not Graded: Vision-Language Models Fail at Slant-from-Texture Perception

Qian Zhang[1], Michal Golovanevsky[1,3], Fulvio Domini[2], and James Tompkin[1]

[1] Brown University Computer Science
[2] Brown University Cognitive Science
[3] Harvard University

**Abstract.** Human perception of surface slant from texture exhibits systematic, graded biases that emerge reliably in psychophysical experiments. Prior work showed that unsupervised CNNs reproduce several human-like biases, while supervised CNNs do not. Do Vision-Language Models (VLMs) exhibit similar competences? Across multiple VLM families and model scales, zero-shot and in-context prompting both produce distinctive failures: slant is predicted at only a small set of anchors (e.g., 0°, ±25°, ±45°) with little dependence on stimulus field of view, optical slant, or surface curvature. Supervised fine-tuning partially remediates the failure, but residual anchoring persists. While success in high-level vision-language benchmarks might not require sensitivity to low-level geometric cues, we interpret anchoring as a failure at the representation-to-output language interface: not necessarily an absence of geometric encoding, but a failure to express it in a graded form.

## 1 Introduction

Comparing human and neural network visual systems has a long anecdotal history, but recent large models trained on billions of images have made the comparison of their behaviors and mechanisms more meaningful. Understanding these relations through *in silico* experiments is an important direction for vision science, e.g., to raise hypotheses for how biological vision might function. But it is also useful for practical tasks. Often, human-AI collaboration is predicated on AI systems being able to predict what humans perceive (AI: *"I can see that."* Human: *"I cannot see that."*). Yet, studies of whether the basic psychophysical abilities of AI systems match those of humans are rarer, as most works focus on downstream performance.

Take the simple task of estimating the slant of a surface. Humans are able to judge three-dimensional surface slant from texture gradients alone, yet these judgments are systematically biased rather than accurate. Psychophysical experiments using textured slant surfaces have identified at least three consistent biases in perception [41]: (B1) convex surfaces appear steeper than concave surfaces of equal physical slant; (B2) larger fields of view produce greater perceived slant; (B3) curvature-sign discrimination degrades at small fields of view, falling to chance at 5°FOV. These biases are commonly quantified by perceptual gain: the ratio of judged slant to ground-truth slant. This is approximately 0.56 for regular dot textures, indicating substantial underestimation of surface slant. Curvature-sign accuracy for such textures

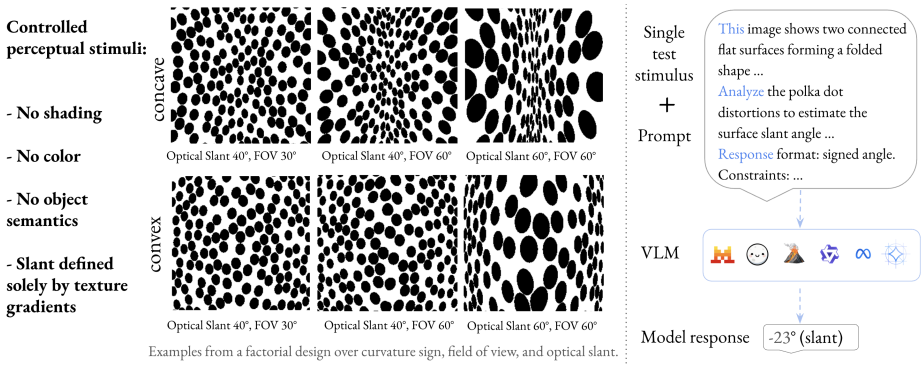



**Fig. 1: Slant-from-texture as a controlled perceptual task for VLMs.** *Left:* Synthetic dot-textured surfaces vary in curvature sign, field of view, and optical slant. Texture gradients are the only cue to 3D orientation. *Right:* Given a single image and instruction, we ask VLMs to estimate surface slant and curvature sign from texture alone.

averages around 86% overall but drops sharply at narrow fields of view. Thus, human slant-from-texture perception reflects systematic image-level heuristics, such as sensitivity to texture scaling gradients, rather than recovery of precise surface geometry.

Do neural networks exhibit similar biases? Wang et al. [47] investigated this question using convolutional neural networks (CNNs) trained on synthetic dot-textured stimuli. Unsupervised CNN autoencoders trained with a reconstruction objective reproduced human-like slant-from-texture biases. Human-like error patterns were recovered from the CNN's internal representations using simple linear projections, which links representational structure to perceptual behavior. However, CNNs trained with direct supervision to regress physical slant achieved unbiased performance. This dissociation indicates that human-like biases arise not from architecture alone, but from learning objectives that emphasize texture statistics rather than explicit geometric labels. These results raise the question of whether such biases persist in visual representations both as models move beyond convolutional architectures and as they are trained under fundamentally different objectives.

Vision-Language Models (VLMs) present such a case. These models combine large-scale vision encoders, such as Vision Transformers, pretrained on broad image distributions, with language models trained on large-scale text corpora. Each is aligned through multimodal supervision on image-caption pairs. The architectural differences from CNNs and the distinct training routine is often assumed to yield representations that are richer or more semantically structured [9,34,53,54]. Recent work has begun to question the geometric ability of VLMs in other domains [15,38], but no studies have examined texture-based slant perception. Such stimuli contain no objects or semantic content; they contain only texture gradients that must be interpreted geometrically. Accordingly, despite strong performance on a range of vision-language benchmarks [3, 17, 21], it remains unknown whether VLMs exploit low-level texture cues in a manner comparable to human perception or unsupervised vision models and whether they can communicate slant in language.

We evaluate multiple VLM families on a classical slant-from-texture task, using stimuli and ground truth matched to prior human and CNN studies. We ask two

questions. First, behavioral: do VLMs exhibit the same systematic biases as humans and unsupervised CNNs, e.g., convex-concave asymmetry (B1) and field-of-view effects (B2, B3). We find that VLMs in zero-shot settings exhibit strong *anchoring* to a small set of discrete values, with predictions failing to exhibit monotonic dependence on stimulus parameters such as optical slant, field of view, or curvature sign—unlike the systematic biases observed in humans and unsupervised CNNs. Second, intervention: can standard adaptation techniques—prompt engineering, in-context learning with labeled examples, or supervised fine-tuning—induce such biases? We show that prompt engineering and in-context learning do not help. Supervised fine-tuning does introduce a monotonic relationship between predicted and ground-truth slant and improves curvature-sign discrimination, but does not eliminate anchoring: predictions remain clustered at discrete values rather than varying continuously. Probing the vision module within four VLMs with different architectures reveals a strong correlation between features and geometric quantities, suggesting that response anchoring is a language "readout" problem. Together, these findings establish slant-from-texture as a *diagnostic* task for revealing systematic differences in how humans, unsupervised vision models, and language-supervised multimodal models respond to texture-based geometric cues, and show where effort is needed to produce future models that can predict (as needed) the human response to the world around us.

*Assumptions and limitations.* 1) Our experiments cover slant from random polka dot textures only, rather than a broader space of textures that includes regularity and shape variation [40] or natural image textures. As psychophysical studies have shown a strong response in humans and CNNs to our polka dot slant stimuli, they are useful probes to reveal existing VLM limitations. 2) We evaluate a breadth of models to characterize the landscape of VLMs, but specific alternatives may vary in significant ways. 3) Our findings do not obviate the usefulness of VLMs in end tasks or imply that VLMs must operate like humans; instead, they show differences in current VLM predictions with respect to human judgment that should be known when applying VLMs to tasks that require similar judgments.

## 2 Dataset and Experiment Setup

*Stimuli.* We produce synthetic dot-textured surfaces following Todd et al. [41, 42], and as followed by Wang et al. [47] (Fig. 1). These have no shading, color, or silhouette cues—slant must be perceived from texture gradients. Each stimulus depicts two textured flat surfaces forming a dihedral angle, varying along four parameters:

| | |
|---|---|
| Optical slant $\sigma_{\text{cen}}$ | 10 values from 25° to 60° |
| Field of view (FOV) | 10 values from 5° to 60° |
| Curvature sign | Convex or concave |
| Texture | 12 i.i.d. random dot jitters (with different random seeds) |

The surfaces have a true physical slant $\rho$ that is their angle relative to the fronto-parallel plane. It is derived from $\sigma_{cen}$ and FOV to preserve perceptual uniformity:

$$\rho = \sigma_{\text{cen}} - s \cdot \frac{\text{FOV}}{4}, \quad s = -1 \text{ (concave)}, \quad s = +1 \text{ (convex)}.$$

Optical slant $\sigma_{\text{cen}}$ is the controlled experimental parameter matched across curvature conditions; physical slant $\rho$ is what observers judge. Optical slant corresponds directly to the texture gradients while physical slant requires integrating curvature and FOV information. This makes physical slant prediction require geometric understanding of 3D structure rather than a direct "readout" of local texture cues.

The first three attributes form a 200-condition factorial design (10 slant $\times$10 FOV $\times$2 curvature). As each instantiation of the 200 conditions has a random texture, each stimulus provides a new projection of the same geometry. Geometric reasoning requires an understanding of distortions to slant generalizable over appearance noise from texture jitter. We split the stimuli into two sets: 2000 for any training (fine-tuning, in-context learning, etc.) and 400 for testing.

*Prompts.* Each stimulus is paired with a text prompt. We vary the information provided in the prompt and its style to try to avoid any particular pitfall in how we describe the problem in language. This design allows us to test whether, and how, linguistic framing modulates visual estimations. The prompt varies the task, whether additional cues are given (e.g., *"The greater the slant angle, the more distorted the dots will appear."*), and the style of prompt in its input and output. This produces seven key variants for physical slant angle prediction and three for binary concave/convex prediction (Tab. 1). We state all prompts in the supplementary material.

| | |
|---|---|
| Task type | Slant prediction (regression) vs. curvature sign (binary classification). |
| Instruction detail | Minimal vs. enriched with cue descriptions. |
| Language style | Natural (colloquial, e.g., "ridge/valley", "fold") vs. technical (e.g., "concave/convex", "dihedral angle"). |
| Output format | Free text vs. structured JSON. |
| In-context | Absence vs. presence of example stimuli with slant labels. |

Additionally, we provide an in-context learning prompt modifier. Along with the test stimuli, such prompts include four labeled example stimuli from the training set that are balanced across slant angle, field of view, and curvature. Labels are provided as free text, details in the supplementary material.

**Table 1: Task prompt configurations.** Natural language style prompt variants and their component. *Left:* Slant prediction as a regression task. *Right:* Curvature sign as a binary classification task.

| Prompt | Components Included |
|---|---|
| 0 | prompt_minimal |
| 1 | setup + task + format |
| 2 | setup + task + format_eg |
| 3 | setup + task + format_eg_json |
| 4 | setup + task + cues + format |
| 5 | setup + task + cues + format_eg |
| 6 | setup + task + cues + format_eg_json |

| Prompt | Components Included |
|---|---|
| 0 | prompt_minimal |
| 1 | setup + task + format |
| 2 | setup + task + cues + format |

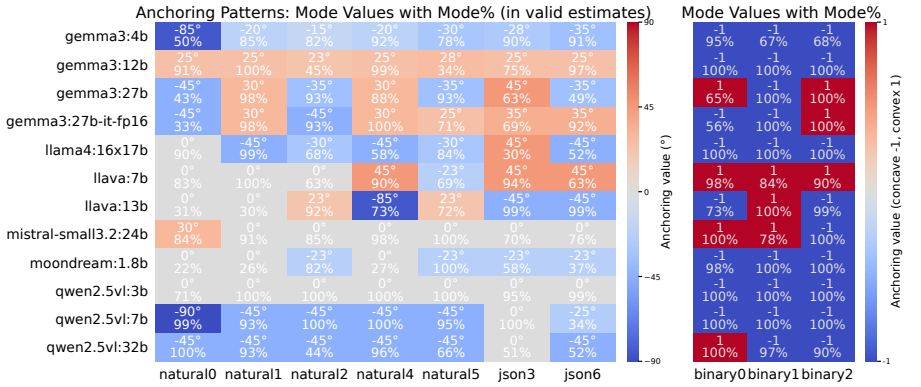



**Fig. 2: VLMs anchor significantly on slant angle and sign prediction.** Anchoring patterns for natural language prompts in regression task (left) and binary classification task (right). The heatmap shows distributions of estimated slant across VLMs (rows) and natural prompt variants (columns). All models collapse to a few discrete anchors (e.g., 0°, ±25°, ±45°) regardless of prompt detail, indicating that prompt variation does not mitigate anchoring. Prompt indices (e.g., "natural0") and their components are in Tab. 1, including both with and without JSON output constraints.

*Model Families and Runs.* We evaluated six recent open-sourced VLMs (Gemma3 [17], LLaMA4 [27], LLaVA1.6 [21], Mistral [16], Moondream [28], Qwen2.5-VL [33]) across multiple model parameter sizes, via local hosting using the Ollama API (v0.11.10) [29]. We use pretrained weights without fine-tuning unless specified. Qwen2.5-VL was also evaluated using HuggingFace Transformers [32] for SFT, probing, and attention-head ablation to access model weights. Additional results from 12 recent VLMs, including both open-source on Hugging Face (e.g., InternVL3.5, Qwen3.6) and frontier closed-sourced models (e.g., GPT-5.4, Gemini3.1, Claude Opus), are provided in the supplementary material.

Each stimuli-prompt pair was submitted in prompt-completion format, and requests were executed asynchronously using multiprocessing. From the responses, we parsed numerical slant estimates or curvature sign classes. We balanced the number of concave and convex test stimuli across experiments. To capture trial-level variance, we repeated each query 10 times at the default temperature ($T = 0.7$). As the observed variance was negligible, subsequent experiments used a lower temperature ($T = 0.1$) with a single run per query for efficiency.

## 3 Results and Analysis

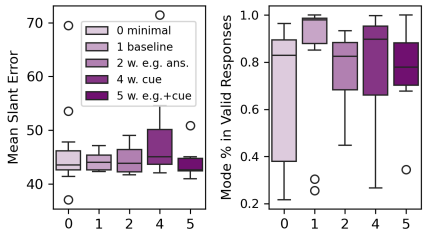
### 3.1 Zero-shot slant estimation

*Systematic anchoring.* Across all models evaluated, zero-shot VLM predictions exhibit large errors in slant and curvature sign (Fig. 2). Rather than varying smoothly with stimulus parameters, predicted slant values cluster around a small set of discrete anchor values, most commonly 0°, ±25°, and ±45°, regardless of prompt detail. The percentages below each cell indicate the mode (most frequent number) as a proportion of all responses. In many cases, the same value appears in more than 50% of responses, indicating anchoring rather than response noise.

We observed three qualitative response patterns:

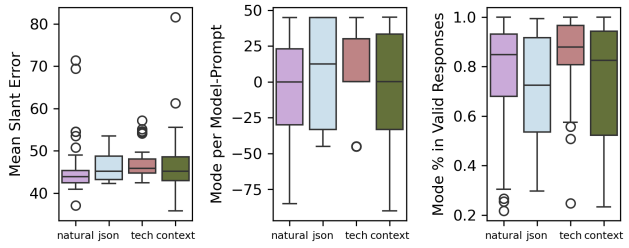
– *Instruction compliance errors:* Models ignored format constraints, outputting text without numerical values or producing out-of-range values.
– *Exemplar anchoring:* Prompts with examples induced "example anchoring" in some models, e.g., Moondream copying example prompt values verbatim ("-23°"), while other models were not affected by the specific values provided.
– *Zero-degree anchoring:* Models predicted flat surfaces (0°), sometimes accompanied by contradictory free-text reasoning in earlier runs (e.g., predicting flat while describing a steep slant).

*Anchoring persists across task framing, cues, prompt styles, and output formats.* We compute the mean absolute error of slant estimates and variance for prompt detail variants, prompt styles and output formats, and task framing (Fig. 3). We observe no significant improvement in VLM ability to predict slant or sign. Varying model temperature did not improve performance. Trial-level consistency analyses at temperature 0.7 showed stable anchoring across repetitions with low variance; averaging predictions did not improve performance. Anchoring is a systematic response pattern rather than stochastic variability.



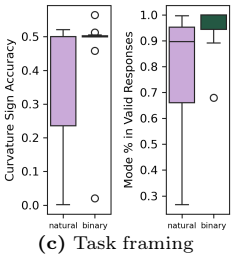

**(a)** Prompt detail (natural style)

Varying prompt detail does not reduce anchoring. The anchor mode value changed for some models, but median errors are above 40° and anchoring dominates. Prompts emphasizing geometric cues yielded no consistent improvement (e.g., descriptions of texture scaling as included in prompts 4 and 5).




**(b)** Prompt style and output format

Prompts as natural vs. technical, or in JSON, or with context show no significant variation (two-way ANOVA on main effect of prompt family on slant error; Tab. 2). JSON formatting did not show consistent reduced example anchoring in any models but did exhibit increased overall variance, i.e., mode percentages decreased.




**(c)** Task framing

Curvature sign prediction as convex vs. concave (classification, "binary" in plot) yielded no improvements over slant prediction (regression, "natural" in plot), with accuracies around 0.5 (chance) and mode percentage increasing to around 90%. This trend is consistent across models.

**Fig. 3:** Effects of prompt and task framing as boxplots of 95% Confidence Intervals.

| Source | sum_sq | df | F | p-value |
|---|---|---|---|---|
| C(model) | 753.19 | 6 | 4.81 | 0.000291 |
| C(prompt_type) | 139.32 | 3 | 1.78 | 0.157403 |
| Interaction | 1230.32 | 18 | 2.62 | 0.001634 |
| Residual | 2192.87 | 84 | – | – |

**Table 2:** Two-way ANOVA (with interaction) of slant error mean for different models and prompts. Prompt type alone does not have a significant effect but depends on the model, e.g., one model might vary with JSON while another does not. Overall, there is no consistent main effect.

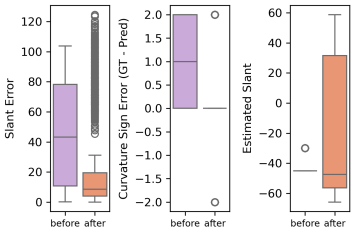



**Fig. 4: Fine tuning only partially helps.** Qwen2.5-VL before and after supervised fine-tuning: slant error decreases and anchoring weakens after SFT, but many outliers persist with high error.

*In-context learning does not help.* In-context prompts do not significantly differ from other prompt families (Fig. 3b): Median slant errors remain above 40°, curvature-sign accuracy stays near chance, and mode percentages remain high. ANOVA revealed no significant main effect of prompt type ($p$=0.157; Tab. 2). While specific anchor values sometimes shift across prompt conditions, the overall anchoring pattern persists.

*Summary.* In zero-shot settings, VLM predictions collapse to discrete anchors, showing no systematic dependence on stimulus parameters, and are unaffected by prompt manipulations, including when given labeled examples for in-context learning. Model family and size modulate variability but do not alter the underlying anchoring pattern. Model performance is uniformly poor, making differences between models difficult to interpret. This behavior is consistent with evidence that LLM-style decoders exhibit anchoring effects and produce coarse numeric estimates that concentrate on a small set of preferred values (including round-number biases) [24, 37, 44]. A complementary explanation is that round numbers are disproportionately frequent in natural language use, making a few canonical numerals strong attractors when models generate numbers as text [49].

### 3.2 Supervised fine-tuning (SFT)

We fine-tuned Qwen2.5-VL-3B using LoRA with a masked token loss on labeled slant and curvature data, updating the vision-language fusion layers (q_proj and v_proj).[1] Qwen2.5-VL-3B was chosen for its consistent anchoring patterns across regression and classification settings and its moderate size for efficient fine-tuning.

*Improved slant estimation, smaller-scale anchoring, larger errors at low FOVs.* SFT substantially improves slant estimation and curvature-sign discrimination (Fig. 4). Mean slant error decreases, and the distribution of predicted slant values broadens

[1] LoRA was restricted to q/v_proj following standard PEFT practice [26]; the regression-head result in Sec. 3.3 confirms this choice does not determine the bottleneck as frozen post-projector features are already linearly decodable for slant.

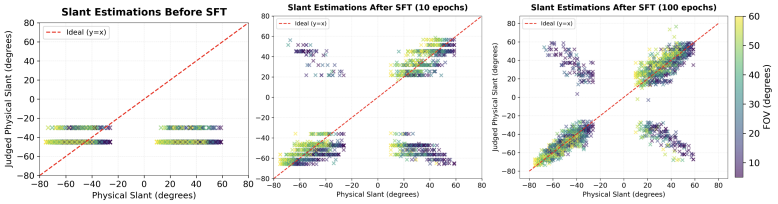



**Fig. 5: SFT improves slant prediction but discrepancies remain.** Judged vs. ground-truth slant angles. From left to right, top to bottom: VLMs before supervised fine-tuning show anchoring; bar chart of slant error distribution; VLMs with 10 epochs of SFT; VLMs with 100 epochs of SFT. After SFT, VLMs show improved slant prediction, but persistently estimate incorrect curvature sign, especially for low FOVs.

beyond the discrete anchors observed before fine-tuning (MAE: $45.1° \rightarrow 15.3°$; STD: $34.9° \rightarrow 26.2°$). Paired statistical tests confirm that fine-tuning yields significant improvement ($df = 1999$, $p = 3.4 \times 10^{-106}$). Curvature-sign accuracy improves from random chance (50%) to 86.10%. It is numerically comparable to human performance (86%) [41] and remains below that of unsupervised CNNs (96.4%) [47], though the underlying error distributions differ qualitatively. Despite these gains, anchoring is not eliminated. Compared to before fine-tuning, anchoring occurs at a finer scale and is distributed around the ground-truth slant (red line), forming horizontal bands near the red line of true prediction (Fig. 5). These decrease as training epochs increase, but remain present.

*Curvature sign remains in error.* We also observe that curvature-sign errors persist after SFT. We analyze curvature sign prediction accuracy as a function of field of view and optical slant in Fig. 7. Shape discrimination errors are elevated at small FOVs and decrease systematically with increasing FOV and optical slant. These errors are *asymmetric* across curvature sign: concave stimuli show near-zero flip rates at large fields of view, whereas convex stimuli retain moderate error rates (0.3–0.4) across optical slant. This pattern varies in the same direction as human behavior (B1), though is more pronounced. The curvature-sign judgement errors change with FOVs, larger at low FOVs (Fig. 5), which is consistent with the FOV dependent error patterns observed in human judgments (B2 and B3). We plot the accuracy of curvature-sign prediction as a function of optical slant and FOV and compare with CNNs results and human study in Fig. 6.

*Summary.* SFT effectively reduces anchoring but does not eliminate it. Predictions become monotonically related to ground-truth slant and curvature-sign discrimination improves and becomes closer to past human and CNN studies, though some slant errors and curvature-dependent asymmetries persist. This suggests that while SFT can steer model behavior, it does not fully override the anchoring pattern.

### 3.3 Probing the Vision Module

Pre-trained VLMs are trained on millions, often billions, of image–caption pairs, from which a complex structure is thought to emerge in their representations. Given that VLMs perform reasonably when directly trained on the stimuli and

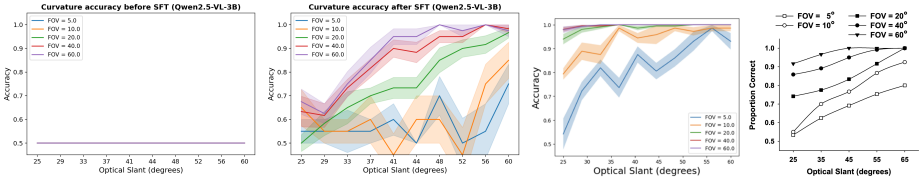



**Fig. 6:** VLM, CNN, and human curvature sign judgment accuracy versus optical slant, shown in matching axes for direct comparison. *Left:* Performance of VLM Qwen2.5-VL (3B) before supervised fine-tuning (SFT). *Center left:* After SFT. Qwen2.5-VL shows improved accuracy with increasing field of view, but still maintains high variance and is overall a poor proxy for human ability. *Center right:* Unsupervised CNN judgments reproduced from Wang et al. [47]. *Right:* Human judgments reproduced from Todd et al. [41].

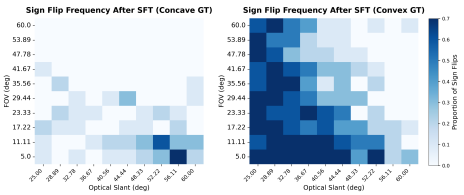



**Fig. 7:** SFT error analysis grid by slant angle and FOV, showing frequency of sign flips in predictions. Larger FOVs and larger optical slant images induced fewer errors (sign flips); no curvature sign errors for concave FOV$>45°$; convex-to-concave misjudge is more frequent than vice versa.

task, we use linear probing [2] to *localize* where correct behavior fails to emerge. The key question is whether zero-shot VLM failures arise from incorrectly encoded geometric information or from a deficit in language readout.

*Probing setting.* We analyze the vision encoder of Qwen2.5-VL-3B, which contains 32 transformer blocks (hidden size 1280, 10 heads) with RoPE and RMSNorm. Images are tokenized by a Conv3D patch embedding ($14\times14$ patches) into 1280D tokens. A subsequent patch merger merges $2\times2$ tokens via a two-layer MLP, reducing token count by $4\times$ and projecting to 2048D. For $224\times224$ stimuli, this produces 81 vision tokens (2048D) passed to the language model. To probe, we extract the 81 post-merger vision tokens in a zero-shot setting (no fine-tuning). Because feature dimensionality (2048) is comparable to sample count (2000), we use ridge regression ($\alpha=1.0$) for all VLM probes to prevent overfitting; all reported $R^2$ values are evaluated on held-out test data.

*Post-merger probing results.* We evaluate regression fits using the coefficient of determination ($R^2$), which measures the proportion of variance in the target variable (PS, OS, FOV) explained by a probe trained on the latent tokens. On the test set, $R^2$=1 indicates perfect prediction, while $R^2$=0 indicates performance equivalent to predicting the mean. Ridge probes on mean-pooled VLM vision tokens achieve physical slant $R^2$=0.826, optical slant $R^2$=0.988, FOV $R^2$=0.884, and curvature accuracy $=0.983$, indicating that geometric information is strongly encoded in the vision representations. These results place the VLM vision encoder in the same performance regime as a pretrained ViT-MAE-86M [14] (FOV $R^2$=0.874, PS 0.856, OS 0.997), suggesting VLM vision features capture geometry at a level comparable to this standalone vision backbone.

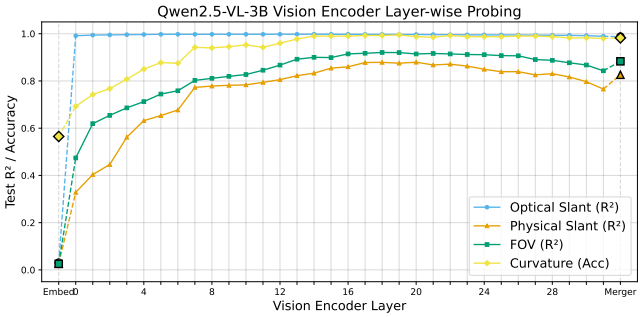



**Fig. 8:** Layer-wise linear probe performance across the 32-layer Qwen2.5-VL-3B vision encoder. Optical slant is near ceiling from the first transformer layer (actual $R^2 = 0.992$–$0.998$; appears flat at this scale). FOV, physical slant, and curvature require progressively deeper processing, peaking around layers 17–19 before declining in the final layers. Embed = Conv3D patch projection output (before any transformer processing). Merger = post-transformer spatial merging (2×2 tokens → 2048-dim). Ridge regression ($\alpha = 1.0$) on mean-pooled 1280-dim features (2048-dim for merger).

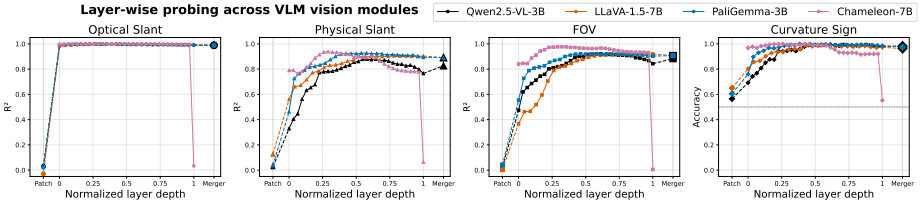



**Fig. 9:** Layerwise probing across four VLMs vision modules, extending the single-model analysis in Qwen2.5-VL-3B to LLaVA-1.5-7B (CLIP-ViT-L/14), PaliGemma-3B (SigLIP-So400M), and Chameleon-7B (VQ-VAE). The result is consistent with the encoding being architecture-general, with the readout bottleneck arising at the language interface rather than in the vision tower.

*Layer-wise probing.* To trace how geometric information develops across the vision encoder, we extract mean-pooled features from every intermediate layer using forward hooks: the patch embedding output, all 32 transformer blocks, and the post-merger output. Applying ridge regression probes independently at each layer reveals a clear progression (Fig. 8). The raw patch embedding contains almost no geometric information (OS $R^2 = 0.031$, all others near zero; curvature accuracy $= 0.565 \approx$ chance), indicating that the Conv3D projection captures little beyond local texture statistics. However, by the first transformer block (layer 0), optical slant is already near ceiling ($R^2 = 0.992$), confirming that it is a low-level image property recoverable from local texture gradients after minimal processing.

In contrast, FOV, physical slant, and curvature require progressively deeper processing: FOV $R^2$ rises from 0.474 (layer 0) to 0.920 (layer 18); physical slant from 0.329 to 0.879; curvature accuracy from 0.690 to 0.995. The correlation of all three variables peaks around layers 17–19 (the 56th–59th percentile of the 32-layer network) and then *declines* in the final layers: by layer 31, FOV drops to 0.844, physical slant to 0.769, and curvature to 0.983. The patch merger partially recovers

**Table 3:** Layer-wise probe for four VLMs with different vision encoders and language models. $R^2$ values listed for each geometric property at the layer where they peak.

| Model | Vision tower | LM | Fusion | OS peak | FOV peak | PS peak | Curv peak |
|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-3B | Qwen-native ViT | Qwen2.5 | Late | 0.998 | 0.921 | 0.879 | 0.995 |
| LLaVA-1.5-7B | CLIP-ViT-L/14 | Vicuna | Late (frozen vision) | 0.997 | 0.923 | 0.905 | 0.990 |
| PaliGemma-3B | SigLIP-So400M | Gemma-2B | Late (prefix) | 0.997 | 0.921 | 0.926 | 1.000 |
| Chameleon-7B | VQ-VAE | unified | Early (vocab fusion) | 0.999 | 0.977 | 0.938 | 1.000 |

from this late decline (FOV = 0.884, PS = 0.826), likely because its spatial 2×2 merging reintroduces local structure that the late layers had redistributed.

This late-layer decline is consistent with the view that the final transformer blocks specialize in the downstream language modeling objective, i.e., reorganizing features for cross-modal projection at the cost of some geometric accessibility. The practical implication is that middle layers of the vision encoder (around layer 18) contain the richest geometric signal, not the final output.

*Generalization across vision encoders.* To test whether the mid-/late-layer geometric-encoding peak is specific to Qwen2.5-VL's vision tower, we replicate the analysis on three additional VLMs with different vision encoders and language models: LLaVA-1.5-7B [1], PaliGemma-3B [4], and Chameleon-7B [39] [2] (Tab. 3). For each, we hook the per-layer outputs of the vision tower (including the projector), mean-pool the patch tokens, and fit ridge regressions (FOV, optical slant, physical slant) and a logistic probe (curvature sign).

As shown in Fig. 9, we found that geometric encoding generalizes across vision-encoder architectures. Across all four models, the layer-wise $R^2$ profile is qualitatively similar: optical slant saturates at $R^2 \geq 0.997$ by the first encoder layer; physical slant climbs gradually to a peak of 0.88–0.93 at 60–80% normalized layer depth; curvature accuracy reaches 0.99–1.00; and the projector preserves $\geq$ 95% of the late-layer peak. Our finding holds across vision towers.

One model of interest is Chameleon: the only model where vision and text tokens flow through the same transformer. Layer 31, the final layer feeding into the language model (LM) head, shows a deep drop in geometric correlation: OS $R^2 = 0.033$, FOV $R^2 = 0.005$, PS $R^2 = 0.065$, curvature accuracy = 0.552 (chance). The geometric information present at layers 12–30 ($R^2 > 0.9$) is transformed away in the final layer toward the language-output distribution. This supports our claim that encoding is preserved deep in the network, but the final language-generation step discards it.

*Readout test.* To test whether this is a true readout problem, we regress on the LM-input Vision Tokens, i.e. the vision tokens projected to LM embedding space contain the necessary geometric information but the language model cannot translate it into a continuous estimate. We mean-pooled the frozen post-projector tokens and trained a single ridge-regression head on the 2000-image training split, with no further tuning. As shown in Fig. 10, on the held-out 400-image test set, this

[2] As an architecture-diversity test, the sizes are each architecture's canonical released size, not a controlled choice: LLaVA-1.5 has no 3B and its smallest is 7B; PaliGemma is 3B only. A true size-controlled SFT comparison would be a separate experiment.

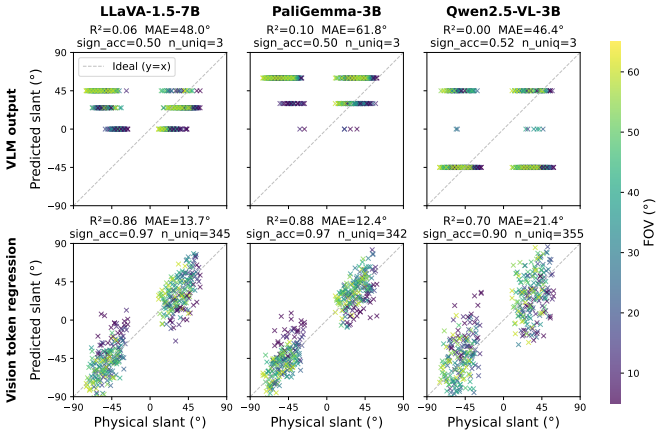



**Fig. 10:** Readout bottleneck—VLM output anchors (top); LM-input vision tokens encode continuous slant (bottom). We extract the mean-pooled post-projector vision tokens (in LM embedding space) from Qwen, LLaVA, and PaliGemma and train a linear regressor on train image tokens to predict physical slant. Results on held-out test images.

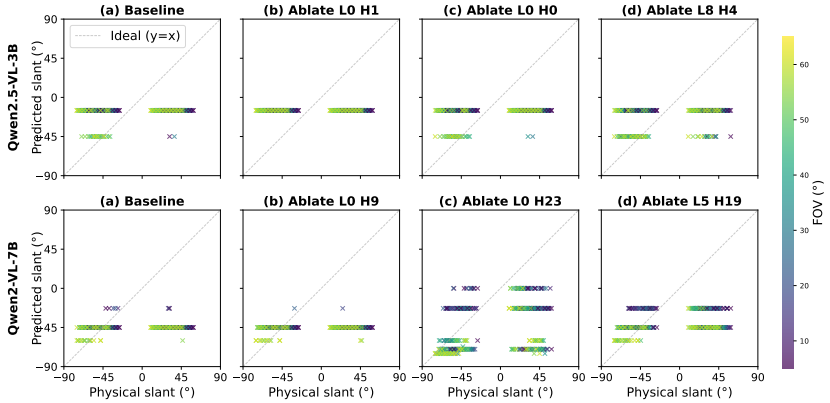



**Fig. 11:** No single attention head ablation produces continuous stimulus-dependent predictions. We apply mean ablation on Qwen2.5-VL-3B (top) and Qwen2-VL-7B (bottom) slant prediction. (a) Baseline predictions; (b, c, d) Typical results of ablating attention heads at different layers. The prediction barely changes; some even induced prompt copying, e.g. ablating L5 H19, a new anchor at "−23°" appears that matches the prompt example.

single linear projection predicts physical slant at $R^2 = 0.696$ (Qwen2.5-VL-3B; MAE 21.4°), $R^2 = 0.857$ (LLaVA-1.5-7B; MAE 13.7°), $R^2 = 0.880$ (PaliGemma-3B; MAE 12.4°), and produces 342+ of 400 unique predicted angles, i.e., predictions are essentially continuous. This shows directly the readout bottleneck, as the geometry is decodable but the language head discards it.

### 3.4 Ablating Language Model Attention Heads

Does intervention reduce anchoring? Recent work by Rudman et al. found individual attention heads in VLMs that copy prompts, and showed that ablating specific

heads can eliminate copying of example values from the prompt [37]. Inspired by this approach, we test whether anchoring in our task arises from specific attention heads in the language model. For each targeted head, we replace its pre-projection activations with the mean activation across all batch and token positions, removing position- and content-specific information while preserving activation scale.

We sweep all layer-head pairs in Qwen2.5-VL-3B (36 layers $\times$ 14 heads $=$ 504 ablations), ablating one head at a time on the 400-stimuli test set. No single head ablation eliminates anchoring (Fig. 11). Some ablations modestly reduce the dominance of the primary anchor value and increase response variance, but most simply redistribute responses across existing anchor values without producing graded estimates that vary continuously with stimulus parameters.

We repeat the same analysis on the larger Qwen2-VL-7B model [31] (28 layers $\times$ 28 heads $=$ 784 ablations) and observe a similar pattern: individual head ablations slightly redistribute responses but do not substantially reduce anchoring. These results suggest that anchoring in continuous perceptual estimation is not attributable to a single attention head but may instead reflect a distributed property of the language model. Although probing shows that geometric information is encoded in the vision representations, the language model cannot reliably read out continuous slant values. Simple attention-head interventions do not recover graded predictions.

## 4 Discussion

*Anchoring as a language readout failure.* Our central finding is not that VLMs perform poorly on slant from texture, but that they fail in a specific, structured way: predictions collapse to a small set of discrete values regardless of stimulus parameters, rather than scattering around the ground truth. This pattern is qualitatively distinct from humans and unsupervised CNNs, which map continuous stimulus variation to smooth, monotonic changes in perceived slant in consistent biased manner (curvature sign asymmetry B1, FOV dependency B2 and B3). The failure is therefore one of graded expression, not magnitude or precision. Our probing results (Sec. 3.3) localize this failure: geometric information is present and linearly decodable throughout the vision encoder and in the post-merger tokens passed to the language model, yet the language output does not reflect it. In effect, *the model knows the geometry but does not say it.* Directly tracing how this information transforms through the language model's hidden states remains future work.

*Speculative: The cross-entropy bottleneck.* A deeper question is whether the encoding-readout dissociation reflects a mismatch between training objectives. VLMs are trained with next-token prediction (cross-entropy over discrete text tokens): a continuous value like 42.5° must be expressed as a sequence of text tokens ("4", "2", ".", "5"), and the model must learn that this token sequence corresponds to a specific point on a continuous scale. There is no direct encouragement for the model to produce text outputs that co-vary smoothly with the continuous geometric variables encoded in vision tokens.

*Speculative: Anchoring as a cue-deficient default value.* Past human perception research found that distance prediction defaults to a specific scaling value when

distance cues are not present [13]. Our stimuli are deprived of any cues beyond texture gradients, and the models, trained on cue-rich images, may defer to a small set of "default" values. As shown in the probing results, texture-gradient information is present in the LM-input vision tokens, but the language module does not recover it, so the outputs fall back to the default values as it would under cue deficiency. SFT teaches the model to recover slant from these cue-deficient stimuli, reducing anchoring and improving the monotonicity of predictions. Future work could test this by enriching the stimuli with additional geometric cues, to see whether graded predictions are recovered.

*Low FOV predictions.* Our stimuli are $224 \times 224$ pixels. Image resolution, bit depth, and transformer tokenization block size likely affect the ability of the model to make predictions for low FOV stimuli because, at low FOV, texture gradients become small and so are "hard to see." This relates to ideas of human visual acuity.

*Broader implications.* Our findings suggest a tension between language grounding and continuous geometric expression in slant from texture task. VLMs perform well on high-level vision-language benchmarks that emphasize object recognition, scene understanding, and visual reasoning, yet they fail on a task that requires only the interpretation of low-level texture gradients. Standard VLM benchmarks emphasize categorical or semantic judgments that are naturally suited to discrete text output, and may therefore not reveal difficulties in expressing continuous variables. Slant from texture, grounded in psychophysical research, provides a simple but revealing test case for probing the boundaries of elementary visual understanding in multimodal models, especially when considering tasks in which VLMs must accurately predict what a human is likely to see.

## 5 Related Work

*Human perception of slant from texture.* Slant-from-texture is a classic perceptual problem with well documented human biases. Oruç et al. [30] compared linear perspective (a grid of lines) and texture gradient (diamond-shaped texture elements) cues for 75° slant perception. Observers tend to do better with combined cues than with either cue alone. The results were consistent with a linear combination of estimates from cues. Todd et al. showed that human slant-from-texture perception is systematic and biased [41, 42], identifying consistent effects of field of view and curvature sign (biases B1-B3 in our notation). More recent work further studied the effects of different texture cues in cue-conflicted and cue-consistent conditions. Chen et al. found that texture compression influenced slant settings more than other texture cues, with its influence decreasing with larger field of view (10° vs. 20°) and less regular textures [6]. Among texture types, regular blob (polka-dot) patterns are popular for modeling texture perception [20] and produce superior slant discrimination performance in humans compared to uniform lattices, Voronoi tessellations, plaids, and noise [36]. Unlike plaids or contour textures, they lack explicit perspective lines, making them a purer test of texture-gradient processing. They can also be systematically manipulated to test the effects of gradient compression [6] and gradient disruption [19].

*Computational models of slant-from-texture perception.* CNNs trained on texture statistics reproduce human-like slant biases under unsupervised learning objectives [47], supporting the use of deep networks as computational analogues for cue learning and inference. This connects to classic shape-from-texture formulations that recover surface orientation from texture gradients under assumptions such as homogeneity [18,48], and to recent computer vision models that revisit this problem under realistic settings (unknown textures, nuisance factors), showing texture alone can constrain local surface geometry in the wild [45]. In parallel, modern monocular depth estimation learns strong geometric priors from large-scale data, both supervised and self-supervised [10, 12, 50], but these systems typically entangle multiple cues and do not isolate the specific contribution of texture gradients to surface slant.

*Vision-language models for visual perception.* VLMs couple a pretrained vision encoder with a large language model via a lightweight alignment module, enabling strong performance on broad multimodal understanding and reasoning benchmarks [3, 17, 21, 23, 25, 52]. However, their behavior on low-level perceptual and geometric tasks (e.g., fine-grained spatial relations, orientations, and viewpoint-dependent reasoning) is less well understood; existing spatially focused evaluations reveal substantial gaps [22, 51]. Emerging diagnostics suggest that large vision/VLM models exhibit weaknesses in metric depth and geometry understanding too, motivating targeted probes and RGB-D augmentation [5, 7, 8]. Overall, many training and evaluation approaches emphasize semantic recognition and language generation (captioning/VQA/grounding) rather than geometry.

*Failure modes of VLMs.* Controlled perceptual evaluations of VLMs remain rare. Zhang et al. [56] test five visual illusions (color constancy, color assimilation, color contrast, geometry relativity, and geometrical perspective), but most studies use naturalistic images. More broadly, a growing body of work documents systematic VLM failures, particularly on tasks requiring precise visual grounding and geometric reasoning. VLM failures on geometric primitives have been shown to depend upon spurious correlations rather than image evidence [11, 15, 35, 38, 43, 46, 55]. Mechanistic analyses suggest that some errors stem from over-reliance on language signals, including anchoring to prompt-provided cues and copying textual constraints even when they conflict with visual evidence [37]. However, much of this literature relies on naturalistic images, leaving open the question of how VLMs behave under tightly controlled psychophysical-style stimuli that isolate low-level geometric cues. Zhang et al. [57] study image classification failures of VLMs. Their work provides tangential evidence for "underutilized visual features" in VLMs, considering classification where the output space is already discrete. We study a continuous geometric quantity, and find anchoring to a small set of salient values independent of stimulus parameters, which has no analog in discrete classification.

*Acknowledgements:* QZ, JT thank NSF CAREER 2144956, NSF IIS-2107409, and an Andy van Dam Fellowship. We are grateful to the anonymous reviewers for their thoughtful comments, and we thank Zitian Tang, Zhenyu Zhu, Channy Lim, Mikhail Okunev, Daniel Ritchie, and Thomas Serre for insightful discussions.

## References


1. Llava-v1.5-7b-hf (2023), https://huggingface.co/llava-hf/llava-1.5-7b-hf, accessed: 2026-05-11
2. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644 (2016)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. Beyer*, L., Steiner*, A., Pinto*, A.S., Kolesnikov*, A., Wang*, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., Unterthiner, T., Keysers, D., Koppula, S., Liu, F., Grycner, A., Gritsenko, A., Houlsby, N., Kumar, M., Rong, K., Eisenschlos, J., Kabra, R., Bauer, M., Bošnjak, M., Chen, X., Minderer, M., Voigtlaender, P., Bica, I., Balazevic, I., Puigcerver, J., Papalampidi, P., Henaff, O., Xiong, X., Soricut, R., Harmsen, J., Zhai*, X.: PaliGemma: A versatile 3B VLM for transfer. arXiv preprint arXiv:2407.07726 (2024), https://huggingface.co/google/paligemma2-3b-mix-224, accessed: 2026-05-11
5. Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., Zhao, B.: Spatialbot: Precise spatial understanding with vision language models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 9490–9498. IEEE (2025)
6. Chen, Z., Saunders, J.A.: Multiple texture cues are integrated for perception of 3d slant from texture. Journal of Vision **20**(7), 14–14 (2020)
7. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. Advances in Neural Information Processing Systems **37**, 135062–135093 (2024)
8. Danier, D., Aygun, M., Li, C., Bilen, H., Mac Aodha, O.: Depthcues: Evaluating monocular depth perception in large vision models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20049–20059 (2025)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)
11. Gao, J., Pi, R., Zhang, J., Ye, J., Zhong, W., Wang, Y., Hong, L., Han, J., Xu, H., Li, Z., et al.: G-llava: Solving geometric problem with multi-modal large language model. arXiv preprint arXiv:2312.11370 (2023)
12. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019)
13. Gogel, W.C.: The sensing of retinal size. Vision research **9**(9), 1079–1094 (1969)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022), https://huggingface.co/facebook/vit-mae-base, accessed: 2026-05-11
15. Huang, K.H., Qin, C., Qiu, H., Laban, P., Joty, S., Xiong, C., Wu, C.S.: Why vision language models struggle with visual arithmetic? towards enhanced chart and geometry understanding. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 4830–4843 (2025)

16. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 **10**, 3 (2023), https://ollama.com/library/mistral-small3.2, accessed: 2026-03-04
17. Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 **4** (2025), https://ollama.com/library/gemma3, accessed: 2026-03-04
18. Kanatani, K.i., Chou, T.C.: Shape from texture: General principle. Artificial Intelligence **38**(1), 1–48 (1989)
19. Kemp, J.T., Vishwanath, D., Domini, F.: Sensory uncertainty does not drive perceptual discriminability in 3d vision (2024)
20. Knill, D.C.: Surface orientation from texture: ideal observers, generic observers and the information content of texture cues. Vision research **38**(11), 1655–1682 (1998)
21. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
22. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics **11**, 635–651 (2023)
23. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
24. Lou, J., Sun, Y.: Anchoring bias in large language models: An experimental study. Journal of Computational Social Science **9**(1), 11 (2026)
25. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: International Conference on Learning Representations (ICLR) (2024)
26. Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., Bossan, B., Tietz, M.: PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft (2022), accessed: 2026-03-04
27. Meta: Llama4-16x17b, https://ollama.com/library/llama4, accessed: 2026-03-04
28. Moondream: Moondream-1.8b, https://ollama.com/library/moondream, accessed: 2026-03-04
29. Ollama: Ollama (software and documentation). Website, https://docs.ollama.com/, accessed: 2026-03-04
30. Oruç, I., Maloney, L.T., Landy, M.S.: Weighted linear cue combination with possibly correlated error. Vision research **43**(23), 2451–2468 (2003)
31. Qwen Team: Qwen2-vl, https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct, accessed: 2026-03-04
32. Qwen Team: Qwen2.5-vl, https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct, accessed: 2026-03-04
33. Qwen Team: Qwen2.5-vl-3b, 7b, 32b, https://ollama.com/library/qwen2.5vl, accessed: 2026-03-04
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
35. Rahmanzadehgervi, P., Bolton, L., Taesiri, M.R., Nguyen, A.T.: Vision language models are blind. In: Proceedings of the Asian Conference on Computer Vision. pp. 18–34 (2024)

36. Rosas, P., Wichmann, F.A., Wagemans, J.: Some observations on the effects of slant and texture type on slant-from-texture. Vision Research **44**(13), 1511–1535 (2004)
37. Rudman, W., Golovanevsky, M., Arad, D., Belinkov, Y., Singh, R., Eickhoff, C., Mahowald, K.: Mechanisms of prompt-induced hallucination in vision-language models. arXiv preprint arXiv:2601.05201 (2026)
38. Rudman, W., Golovanevsky, M., Bar, A., Palit, V., LeCun, Y., Eickhoff, C., Singh, R.: Forgotten polygons: Multimodal large language models are shape-blind. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 11983–11998 (2025)
39. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024), `https://huggingface.co/facebook/chameleon-7b`, accessed: 2026-05-11
40. Todd, J.T., Thaler, L.: The perception of 3d shape from texture based on directional width gradients. Journal of vision **10**(5), 17–17 (2010)
41. Todd, J.T., Thaler, L., Dijkstra, T.M.: The effects of field of view on the perception of 3d slant from texture. Vision Research **45**(12), 1501–1517 (2005)
42. Todd, J.T., Thaler, L., Dijkstra, T.M., Koenderink, J.J., Kappers, A.M.: The effects of viewing angle, camera angle, and sign of surface curvature on the perception of three-dimensional shape from texture. Journal of vision **7**(12), 9–9 (2007)
43. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9568–9578 (2024)
44. Tseng, C.Y., Roy, S., Thasin, M., Zhang, D., Effiong, B.: Streetmath: Study of llms' approximation behaviors. arXiv preprint arXiv:2510.25776 (2025)
45. Verbin, D., Zickler, T.: Toward a universal model for shape from texture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 422–430 (2020)
46. Vo, A., Nguyen, K.N., Taesiri, M.R., Dang, V.T., Nguyen, A.T., Kim, D.: Vision language models are biased. arXiv preprint arXiv:2505.23941 (2025)
47. Wang, Y., Zhang, Q., Aubuchon, C., Kemp, J., Domini, F., Tompkin, J.: On human-like biases in convolutional neural networks for the perception of slant from texture. ACM Transactions on Applied Perception **20**(4), 1–18 (2023)
48. Witkin, A.P.: Recovering surface shape and orientation from texture. Artificial intelligence **17**(1-3), 17–45 (1981)
49. Woodin, G., Winter, B., Littlemore, J., Perlman, M., Grieve, J.: Large-scale patterns of number use in spoken and written english. Corpus Linguistics and Linguistic Theory **20**(1), 123–152 (2024)
50. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
51. Yu, S., Chen, Y., Ju, H., Jia, L., Zhang, F., Huang, S., Wu, Y., Cui, R., Ran, B., Zhang, Z., et al.: How far are vlms from visual spatial intelligence? a benchmark-driven perspective. arXiv preprint arXiv:2509.18905 (2025)
52. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024)
53. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. 2022 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18102–18112 (2021)

54. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE transactions on pattern analysis and machine intelligence **46**(8), 5625–5644 (2024)
55. Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Qiao, Y., et al.: Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In: European Conference on Computer Vision. pp. 169–186. Springer (2024)
56. Zhang, Y., Pan, J., Zhou, Y., Pan, R., Chai, J.: Grounding visual illusions in language: Do vision-language models perceive illusions like humans? In: Proceedings of Conference of Empirical Methods in Natural Language Processing. EMNLP 2023 (2023)
57. Zhang, Y., Unell, A., Wang, X., Ghosh, D., Su, Y., Schmidt, L., Yeung-Levy, S.: Why are visually-grounded language models bad at image classification? Advances in Neural Information Processing Systems **37**, 51727–51753 (2024)