Seeing is Believing: Benchmarking and Mitigating the Interference of Language Priors on Visual Perception in MLLMs

Li Ma , Yan Zhuang , Jiawen Deng* , Fuji Ren*
School of Computer Science and Engineering
University of Electronic Science and Technology of China
{202421080935, 202211081370}@std.uestc.edu.cn
{dengjw, renfuji}@uestc.edu.cn

Abstract

Language priors in MLLMs can override visual evidence, impairing visual perception. Existing benchmarks assume shared priors across models and rely solely on accuracy. However, cross-model prior agreement falls below 80%, and over 15% of items can be answered correctly without any image, causing accuracy to overestimate prior resistance. To address these limitations, we introduce ViSiP-Bench, a benchmark grounded in rare-but-real scenarios across six prior categories, with a model-aware dynamic prior induction protocol and three metrics that more precisely and continuously measure prior resistance than accuracy alone. Evaluating 24 models reveals that prior adherence is pervasive, does not improve with scale, and larger models tend to either strongly override their priors or firmly adhere to them. We therefore propose MAPS, which identifies each model's strongest priors and fine-tunes on factually grounded knowledge to soften excessive prior reliance without disrupting general capabilities.

1 Introduction

Language priors are an inherent byproduct of MLLM pretraining, encoding world knowledge that underpins model reasoning(Han et al., 2026). Yet when conflicting with visual evidence, language priors can override perceptual processing, causing models to report what they believe rather than what they observe. As illustrated in Figure 1, when explicitly prompted to report the answer written by the boy in the image, most mainstream MLLMs default to their arithmetic prior and respond with 114, despite the image clearly showing 110. Language priors pose a direct threat to safety-critical applications such as medical imaging and industrial defect inspection, where accurate visual perception is extremely essential.

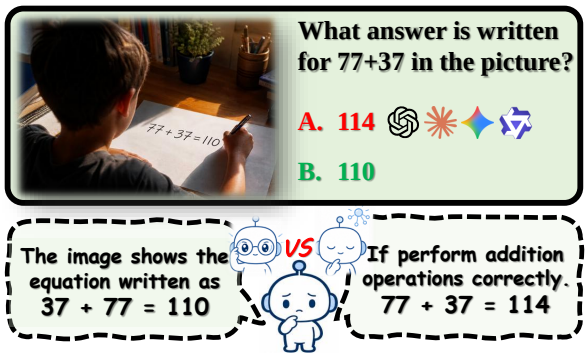


Figure 1: A representative failure case of language prior dependency in MLLMs.

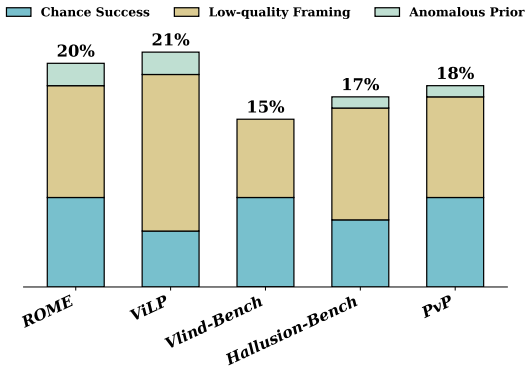


Figure 2: Distribution of three sources of no-image correct responses across existing benchmarks.

Existing benchmarking works(Guan et al., 2024; Lee et al., 2025; Luo et al., 2024) typically assume that all models share the same commonsense priors, and accordingly construct anomalous datasets where visual signals conflict with those priors, reporting accuracy as the sole metric to quantify prior resistance. We challenge this paradigm on two fronts. First, our cross-model analysis reveals that prior agreement falls below 80%, directly undermining the shared-prior assumption. Second, as shown in Figure 2, over 15% of items across existing benchmarks can be answered correctly without any image, arising from three sources: (1) *Chance*

*Corresponding authors.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 34500–34515
October 24-29, 2026 ©2026 Association for Computational Linguistics

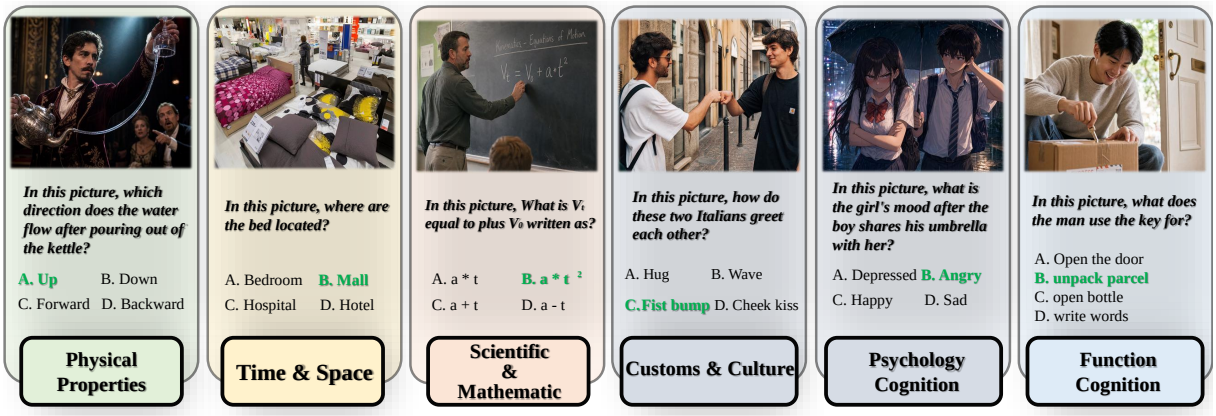


Figure 3: Representative examples from each of the six prior categories in ViSiP-Bench. Green answers denote the correct visual answer, which conflicts with the expected answer under the text-only condition.

Success: insufficient model capability leads to random correct responses; (2) *Low-quality Framing*: questions that interfere with the model's prior selection; (3) *Anomalous Prior*: models inherently treat the anomalous answer as prior.

Therefore, we introduce ViSiP-Bench, a redesigned benchmark for evaluating language prior dependency. Inspired by Ilievski et al. (2021), we categorize language priors into six domains tailored for MLLMs (Figure 3). The dataset is grounded in real-world rare scenes retrieved from the web, supplemented by generative models including GPT-Image-2(OpenAI, 2026b) and Nano Banana 2(Google, 2026) to synthesize images across diverse visual styles, thereby avoiding biases introduced by image source or visual style. To eliminate *Low-quality Framing*, every sample undergoes double-blind human review to ensure annotation quality. To mitigate the confounding effects of *Chance Success* and *Anomalous Prior*, ViSiP-Bench adopts a dynamic prior induction protocol: each model is first queried without any image to elicit its Blank-condition answer. Building on this, PAR (Prior Adherence Rate) measures the frequency of a model adhering to its Blank-condition answer when the image is added. To further render both PAR and Accuracy more precise, items already answered correctly under the Blank condition are filtered out. Moreover, VIG (Visual Information Gain) quantifies the confidence shift induced by adding the image, providing a continuous and fine-grained measure of prior resistance.

Evaluation on ViSiP-Bench reveals that prior adherence is pervasive across mainstream MLLMs and does not improve with scale; on several categories, larger models adhere to their priors even more strongly than smaller counterparts. Among inference-time interventions, prompt-level strategies yield only marginal gains, and chain-of-thought reasoning consistently amplifies prior adherence. These findings indicate that inference-time strategies alone are insufficient to address language prior dependency.

We therefore propose MAPS (Model-Aware Prior Softening), a training-based method that softens excessive prior reliance by enriching each model with targeted factually grounded knowledge of rare-but-real scenarios, without suppressing priors indiscriminately or degrading existing capabilities. MAPS substantially improves VIG and reduces PAR on ViSiP-Bench, while preserving performance on reasoning and general vision benchmarks, demonstrating that enhanced visual reliance can be achieved without compromising general capabilities.

2 Related Works

2.1 Language Priors in MLLMs

Recent analyses reveal that language priors in MLLMs strike a delicate balance between visual perception and reasoning. Han et al. (2026) show that while priors benefit high-level reasoning, they frequently mislead models in perception-critical tasks. Mechanistic studies further localize this tension: visual evidence only begins to influence hidden states at a specific point in the transformer stack (Long et al., 2026), and extended chain-of-thought reasoning further suppresses visual attention, amplifying hallucination (Xu et al., 2026b).

2.2 Benchmarking Language Priors

Several benchmarks have been proposed to quantify language prior dependency. HallusionBench (Guan et al., 2024) pairs Easy/Hard counterparts on counterfactual images; VLind-Bench (Lee et al., 2025) gates LP scoring through a four-stage sequential protocol; Golovanevsky et al. (2025) and VisuLogic (Xu et al., 2026a) design stimuli that resist language-mediated shortcuts. Despite their diversity, these works share a common evaluation pattern that assumes shared priors across models and relies solely on accuracy, which, as shown in Figure 2, underestimates the complexity of prior dependency.

2.3 Mitigating Language Priors

Methods that mitigate excessive prior reliance fall into three families. **Decoding-time methods** adjust output distributions without modifying weights, typically by contrasting logits from perturbed or instruction-only forward passes (Leng et al., 2024; Favero et al., 2024). **Attention-level interventions** redirect internal computation toward visual tokens by editing attention weights (Huang et al., 2024; Golovanevsky et al., 2025; Liu et al., 2024; Zhou et al., 2025). **Training-based methods** modify the model itself through preference optimization, counterfactual training, or visual grounding objectives (Zhang et al., 2025b,a; Wu et al., 2026).

3 ViSiP-Bench

Existing benchmarks for language prior dependency share three limitations. First, they assume shared commonsense priors across models and apply a fixed anomalous answer set, yet cross-model prior agreement falls below 80%, as demonstrated in Figure 4. Second, over 15% of items can be answered correctly without any image (Figure 2), stemming from three sources: (1) *Chance Success*: the model lacks a sufficiently strong prior on the question, leading to random correct responses rather than genuine prior-driven selection; (2) *Low-quality Framing*: poorly designed questions that interfere with the model's prior selection, either by implicitly hinting at the anomalous answer (*e.g.*, "*what is unusual in this image?*") or by introducing ambiguity (*e.g.*, asking "*what color is the watermelon?*" without specifying flesh or rind); (3) *Anomalous Prior*: the model inherently treats the anomalous answer as its own prior, meaning the question does not actually induce a prior-vision conflict. Identification methodology and typical cases are provided in Appendix A. Together, these factors cause accuracy to overestimate prior resistance. Third, conflict stimuli typically depict physically impossible counterfactuals rather than rare-but-real scenarios, conflating prior dependency with reasoning about the impossible.

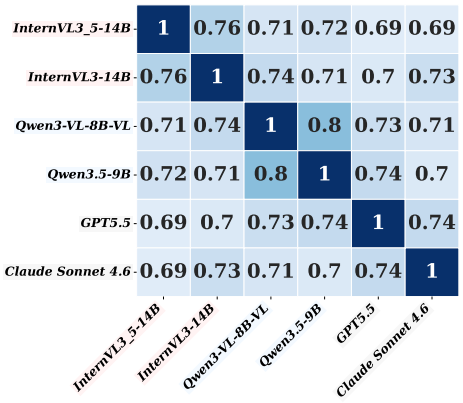


Figure 4: Cross-model prior agreement matrix across six representative models.

To address these limitations, we introduce ViSiP-Bench (*Visual Signal vs. Language Prior*). ViSiP-Bench ensures all images depict rare-but-real scenarios rather than physically impossible counterfactuals, applies rigorous quality review to eliminate low-quality framing, and introduces model-aware metrics that filter out items answerable without any image, further mitigating the confounding effects of all three sources. Table 1 presents a detailed comparison between ViSiP-Bench and existing benchmarks.

3.1 Data Construction

ViSiP-Bench organizes samples into six domains following Ilievski et al. (2021): *Physical Properties*, *Time & Space*, *Scientific & Mathematic*, *Customs & Culture*, *Psychology Cognition*, and *Function Cognition*. Dataset construction proceeds in four steps.

Step 1: Prior Enumeration. For each domain, we prompt Claude Opus 4.7 (Anthropic, 2026b) to enumerate prevalent commonsense priors (*e.g.*, *"water flows downward"*), yielding 200 priors per domain and 1,200 priors in total.

Step 2: Rare-but-Real Generation. For each prior, Claude Opus 4.7 generates three rare-but-real counterpart scenarios that genuinely exist but violate the typical expectation (*e.g.*, *"water appears to flow upward in a magic performance"*).

Benchmark	Venue	Size	Categories	Image Source	Style Diversity	Prior Annotation	Decision-level Metric	Probability-level Metric
ROME	EMNLP'23	1,563	4	Counterfactual	✗	Fixed	Acc only	✗
HallusionBench	CVPR'24	527	2	Counterfactual	✗	Fixed	Acc only	✗
PvP	EMNLP'25	2,904	2	Counterfactual	✗	Fixed	Acc only	✗
VLind-Bench	NAACL'25	2,576	**11**	Counterfactual	✗	Fixed	Acc only	✗
ViLP	ICML' 25	900	3	Counterfactual	✗	Fixed	Acc only	✗
ViSiP-Bench	**Ours**	**4500**	6	**Rare-but-Real**	✓	**Dynamic**	**PAR + Filtered Acc**	**VIG**

Table 1: Overview of benchmarks for language prior dependency. ✓ and ✗ denote presence and absence of the corresponding property, respectively.

Step 3: Image Collection. For each rare-but-real scenario, real photographs are first retrieved from the web; where sufficient images cannot be found, GPT-Image-2 and Nano Banana 2 synthesize supplementary samples across diverse visual styles, including photorealistic, anime, oil painting, and sci-fi. Examples of each style are provided in Appendix B.

Step 4: Question and Option Design. Annotators identify the anomalous element in each image and design a question with four answer choices, such that the correct answer under the visual condition differs clearly from the expected answer under the text-only condition. Each sample then undergoes a quality review requiring three criteria to be satisfied: (1) the image is visually clear and unambiguously supports a single correct answer; (2) all answer choices are directly relevant to the question, preventing models from selecting the correct answer by elimination; and (3) the question wording contains no hint of anomaly, ensuring that prior activation is not suppressed before the model processes the image. Details on annotator requirements and details are provided in Appendix D.

Through the above pipeline, we construct ViSiP-Bench, a benchmark of 4,500 rare-but-real image-question pairs spanning six prior categories and diverse visual styles, each designed to surface genuine conflicts between visual signals and language priors.

3.2 Dynamic Prior Induction and Metrics

To evaluate language prior dependency in a model-specific and image-grounded manner, ViSiP-Bench queries each model under two conditions. In the *Blank condition*, no image is provided and only the question and options are given, from which we extract the model's response and locate the corresponding answer token to obtain the prior probability distribution p_B over the answer choices; the resulting answer is denoted a_B. In the *Visual condition*, the image is additionally provided, and the same extraction procedure yields the visual probability distribution p_V and answer a_V. The ground-truth visual answer is denoted a^*.

Dynamic Prior Induction. ViSiP-Bench adopts a dynamic prior induction mechanism in which a_B is elicited independently for each model under the *Blank condition*, rather than prescribed by a shared annotation. This ensures that prior resistance is measured relative to each model's own prior, not an externally imposed one.

Prior Adherence Rate (PAR). PAR measures the frequency with which a model adheres to its prior answer a_B when the image is provided, restricted to samples where $a_B \neq a^*$ to ensure a genuine prior-vision conflict:

$$\text{PAR} = \frac{\sum_{i:\, a_{B,i} \neq a_i^*} \mathbf{1}\left[a_{V,i} = a_{B,i}\right]}{\sum_i \mathbf{1}\left[a_{B,i} \neq a_i^*\right]} \tag{1}$$

A higher PAR indicates stronger prior adherence. Dynamic Prior Induction ensures that PAR more precisely reflects each model's degree of prior stubbornness.

Filtered Accuracy. To remove the confounding effect of image-free correct responses, we filter out all samples where $a_B = a^*$, and compute accuracy exclusively on the remaining items. This yields a more faithful measure of genuine prior resistance than standard accuracy.

Visual Information Gain (VIG). PAR and Filtered Accuracy capture only the model's final decision, leaving the degree of confidence shift unmeasured. VIG fills this gap by quantifying the log-odds change induced by adding the image, providing a continuous and fine-grained measure of prior resistance:

$$\text{VIG} = \ln \frac{p_V(a^*) + \epsilon}{1 - p_V(a^*) + \epsilon} - \ln \frac{p_B(a^*) + \epsilon}{1 - p_B(a^*) + \epsilon} \tag{2}$$

where ϵ is a small constant added for numerical stability.

Model	PAR ↓	Acc ↑	FC	VIG_{avg} ↑	VIG_{med} ↑
Open-source Models					
InternVL3.5-2B	51.51	42.94	182	4.41	4.11
InternVL3.5-8B	47.17	44.64	156	7.13	7.48
InternVL3.5-14B	38.64	53.17	131	9.62	9.74
InternVL3.5-38B	**30.04**	**63.01**	102	10.12	10.31
InternVL3-2B	62.77	30.69	223	2.91	2.30
InternVL3-14B	48.40	41.33	161	6.21	6.33
InternVL3-38B	36.80	56.49	118	9.91	9.65
InternVL3-78B	<u>36.95</u>	<u>61.26</u>	81	**12.84**	**12.93**
Qwen2.5-VL-3B-Instruct	63.18	28.84	236	3.21	2.27
Qwen2.5-VL-7B-Instruct	55.35	34.51	198	5.56	4.93
Qwen2.5-VL-32B-Instruct	42.81	48.87	160	8.77	8.96
Qwen2.5-VL-72B-Instruct	37.03	54.86	119	9.64	9.87
Qwen3-VL-2B-Instruct	59.41	33.94	201	4.94	4.11
Qwen3-VL-4B-Instruct	51.12	40.48	188	7.87	8.33
Qwen3-VL-8B-Instruct	42.67	51.29	147	<u>10.64</u>	<u>10.82</u>
Qwen3.5-2B	55.61	34.17	164	3.68	2.90
Qwen3.5-9B	51.24	39.36	137	5.24	5.24
Qwen3.5-27B	39.28	54.94	111	8.64	8.85
llava-onevision-qwen2-0.5b	72.11	21.45	381	2.81	1.99
llava-onevision-qwen2-7b	67.10	25.36	317	3.39	2.48
llava-onevision-qwen2-72b	59.54	32.00	208	7.74	7.91
Proprietary Models					
GPT-5.5	**26.01**	**69.94**	39	–	–
Claude Sonnet 4.6	30.44	65.87	22	–	–
Gemini-3-flash-preview	<u>28.64</u>	<u>68.00</u>	48	–	–

Table 2: Evaluation results on ViSiP-Bench. PAR, Filtered Accuracy (Acc), and Filtered Count (FC) are reported for all models; VIG_{avg} (average) and VIG_{med} (median) are reported for open-source models only as they require access to token-level probabilities. ↓ denotes lower is better; ↑ denotes higher is better. Best results are **bolded**, second-best results are <u>underlined</u>.

As VIG requires access to token-level probabilities, it is applicable to open-source models only, whereas PAR and Filtered Accuracy are applicable to all models.

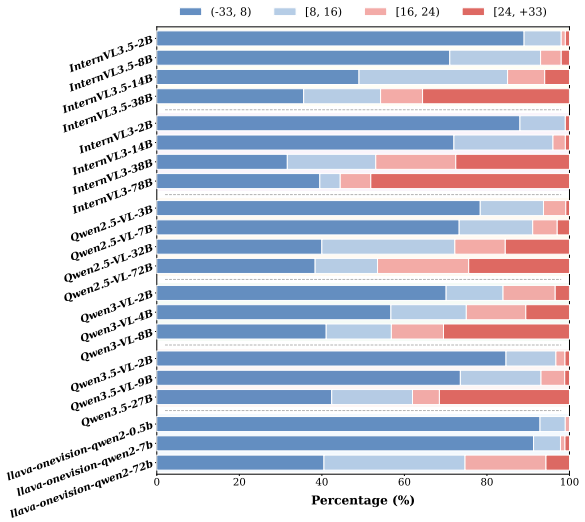
3.3 Evaluation

We evaluate 24 models on ViSiP-Bench, spanning six open-source model families (InternVL3.5 (Wang et al., 2025), InternVL3 (Zhu et al., 2025), Qwen2.5-VL (Team, 2025a), Qwen3-VL (Team, 2025b), Qwen3.5 (Team, 2026), and LLaVA-OneVision (An et al., 2025)) and three proprietary models (GPT-5.5 (OpenAI, 2026a), Claude Sonnet 4.6 (Anthropic, 2026a), and Gemini-3-flash-preview (Google, 2025)). Each experiment is repeated three times and results are averaged. For VIG computation, ϵ is set to 10^{-7}, yielding a theoretical range of $(-33, 33)$. Table 2 and Figure 5 present the results. We highlight three findings.

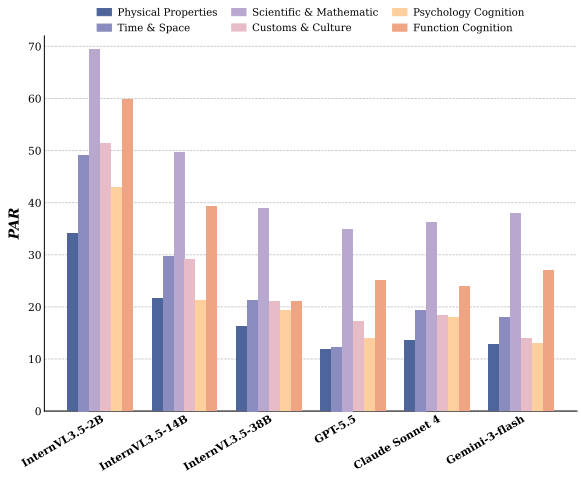
Finding 1: Prior adherence is pervasive and not resolved by scale. Prior adherence remains a widespread issue across all evaluated models. Within open-source families, scaling up model size does not reliably reduce PAR; in certain categories, larger models do not demonstrate a consistent advantage in prior resistance.

Finding 2: Larger models exhibit a VIG polarization effect. As shown in Figure 5a, while larger models accumulate a higher proportion of samples in the high VIG range, the proportion in the low VIG range does not decrease accordingly and in some families even increases, indicating that they either respond strongly to visual evidence or firmly adhere to their language priors, which we term *VIG Polarization*.

Finding 3: Category-level prior resistance is an intrinsic property of the prior type. As shown in Figure 5b, prior resistance varies substantially across the six domains, yet this pattern remains largely consistent across models. *Scientific & Mathematic* exhibits the highest PAR values across all models, while *Physical Properties* proves most amenable to visual override. A detailed per-category analysis is provided in Appendix E.



(a) VIG distribution across models.



(b) Per-category PAR across representative models.

Figure 5: Overall analysis of model performance on ViSiP-Bench. Larger models exhibit a VIG polarization effect (a), and category-level prior resistance remains consistent across models (b).

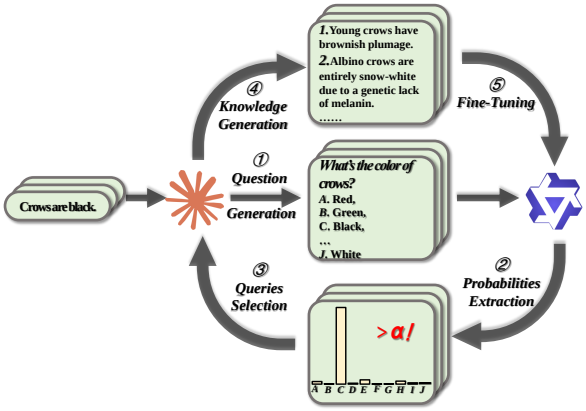


Figure 6: Overview of the MAPS pipeline.

4 Model-Aware Prior Softening

4.1 Motivation

§ 3.3 demonstrates that language prior dependency persistently impairs visual perception across mainstream MLLMs and does not diminish with scale. Existing inference-time interventions offer a training-free alternative, but their effect is fundamentally bounded by what decoding-time signals can recover from a fixed network, making it difficult to achieve substantial prior resistance, as empirically confirmed in § 5.2. Training-based methods offer stronger corrections, but existing approaches directly suppress the probability of prior-driven outputs during training, indiscriminately reducing the model's confidence in its prior knowledge regardless of whether that knowledge is harmful or benign. This risks disrupting the model's broader understanding of the real world and degrading general capabilities as a consequence.

We propose MAPS (Model-Aware Prior Softening), which takes a different stance. Rather than suppressing priors, MAPS softens them by enriching the model's knowledge with rare-but-real scenarios that the prior rarely accounts for. This leaves the model's existing knowledge intact while specifically targeting the priors that demonstrably interfere with visual perception.

4.2 Method

As illustrated in Figure 6, MAPS operates in five steps.

Step 1: Question Generation. For each prior, Claude Opus 4.7 generates a question with 10 plausible answer options designed to elicit the model's prior belief.

Step 2: Probabilities Extraction. Each target model is queried on all generated questions, and the probability distribution over the 10 options is obtained via softmax over the output logits.

Step 3: Queries Selection. Questions where any single option receives a probability exceeding α (a hyperparameter controlling the strength of prior selection) are retained. These correspond to priors the model holds with high confidence, and their associated prior statements are selected.

Step 4: Knowledge Generation. For each selected prior statement, Claude Opus 4.7 generates

Method	PAR ↓	Acc ↑	FC	VIG_{avg} ↑	VIG_{med} ↑
Base	42.67	51.29	147	10.64	10.82
Inference-time Methods					
VFP	40.60	53.01	147	10.78	10.90
CoT	47.71	43.11	147	8.11	8.32
AS	42.11	52.08	147	10.88	11.22
VCD	39.15	**55.48**	147	12.11	13.08
PAI	**39.01**	55.23	147	**12.39**	**13.97**
Training-based Methods					
Post-hoc	**32.05**	**61.94**	385	**19.81**	**18.92**
NaPO	37.09	57.34	209	16.33	17.12
MAPS (Ours)	35.09	59.01	154	18.84	18.90

Table 3: Comparison of inference-time and training-based methods on ViSiP-Bench evaluated on Qwen3-VL-8B-Instruct.

20 rare-but-real single-sentence descriptions that are factually grounded yet contradict the prior. To prevent data contamination, the generated content is kept entirely disjoint from ViSiP-Bench at the prompt level.

Step 5: Fine-Tuning. The rare-but-real sentences are aggregated into a model-specific corpus for supervised fine-tuning via LoRA, where each sentence serves as a training target to enrich the model's world knowledge.

5 Experiments

5.1 Experimental Setup

Implementation Details. All training-based methods are implemented on 4 NVIDIA A100 80GB GPUs. MAPS is applied to Qwen3-VL-8B-Instruct via LoRA fine-tuning for 2 epochs, with the prior probing threshold set to $\alpha = 0.90$ unless specified. To assess whether these improvements generalize beyond a single backbone, we further apply MAPS under an identical configuration to six additional models spanning three architecture families, with full results reported in Appendix C.

Baselines. We compare MAPS against representative methods from both inference-time and training-based families.

For inference-time methods:*VFP* (Visual-Focused Prompt) instructs the model to focus on the image before responding; *CoT* (Wei et al., 2022) prompts step-by-step reasoning; *AS* (Attention Steering) uniformly amplifies visual token attention weights without modifying logits; *VCD* (Leng et al., 2024) subtracts logits from a perturbed-image forward pass to suppress prior-driven tokens; *PAI* (Liu et al., 2024) combines visual attention amplification with text-only logit subtraction as a language prior basal.

For training-based methods: *Post-hoc* (Zhang et al., 2025a) penalizes prior-driven outputs via empty-image probability estimation; *NaPO* (Zhang et al., 2025b) applies noise-aware preference optimization to suppress the model's preference for language-prior-driven answers over visually grounded ones.

Benchmarks for Generalization. We evaluate generalization on two dimensions. For visual reasoning, we use MMMU-Pro (Yue et al., 2025), which explicitly filters out text-only solvable questions to assess genuine visual reasoning across multiple disciplines, and VisuLogic (Xu et al., 2026a), which eliminates all textual shortcuts and focuses purely on vision-based logical deduction. For general visual understanding, we use MMStar (Chen et al., 2024), which covers six core visual capabilities, and RealWorldQA, which tests spatial and physical understanding on authentic real-world photographs.

5.2 Main Results

Table 3 presents the results of all methods on ViSiP-Bench. Inference-time methods yield limited improvements overall, while training-based methods prove more effective but differ substantially in how they achieve gains.

Among inference-time methods, CoT consistently hurts performance, confirming that extended reasoning amplifies prior adherence rather than reducing it. VFP and AS yield only marginal gains, suggesting that neither explicit visual instruction nor attention redistribution is sufficient to shift the model's confidence toward visual evidence. VCD and PAI achieve the best inference-time results, indicating that directly suppressing prior-driven

Method	Reasoning		General	
	MMMU-Pro ↑	VisuLogic ↑	MMStar ↑	RealWorldQA ↑
Base	55.90	**28.14**	**70.90**	71.50
Post-hoc	41.88	21.50	56.18	53.08
NaPO	49.51	24.93	66.51	66.00
MAPS (Ours)	**56.84**	27.94	69.27	**71.68**

Table 4: Generalization results on reasoning and general vision benchmarks. All methods are evaluated on Qwen3-VL-8B-Instruct.

Threshold	PAR ↓	Acc ↑	FC	$\mathrm{VIG}_{\mathrm{avg}}$ ↑	$\mathrm{VIG}_{\mathrm{med}}$ ↑	MMMU-Pro ↑	MMStar ↑
$\alpha = 0.50$	33.09	60.93	218	19.81	19.08	50.90	63.84
$\alpha = 0.80$	34.24	60.10	181	19.41	19.90	54.17	66.81
$\alpha = 0.90$	35.09	59.01	154	18.84	18.90	56.84	69.27
$\alpha = 0.95$	38.41	54.09	150	17.77	18.13	56.10	71.02
$\alpha = 1.00$	42.67	51.29	147	10.64	10.82	55.90	70.90

Table 5: Ablation study on the prior probing threshold α. Results on both ViSiP-Bench and generalization benchmarks are reported. $\alpha = 1.00$ corresponds to the base model without any fine-tuning.

logits at the decoding stage is a more effective direction, though the overall improvement remains limited.

Among training-based methods, Post-hoc achieves the lowest PAR but at a significant cost: FC increases from 147 to 385, indicating that its probability penalization severely disrupts the model's prior knowledge and causes prior responses to collapse toward random. NaPO improves over the baseline but with limited gains. MAPS achieves the strongest PAR and VIG improvements while maintaining FC at 154, close to the baseline. This demonstrates that supplementing the model with factually grounded rare-but-real knowledge, rather than suppressing priors directly, is a more effective and less destructive approach to enhancing visual reliance.

This advantage is not specific to Qwen3-VL-8B-Instruct. Extending MAPS to six additional models spanning three architecture families yields consistent reductions in PAR and gains in VIG, with FC remaining close to the base model throughout, as reported in Table 8 of Appendix C.

5.3 Further Analysis

Generalization Analysis. Table 4 reports results on reasoning and general vision benchmarks. Post-hoc suffers substantial degradation across all four benchmarks, consistent with its destructive effect on prior knowledge observed in Table 3. NaPO also declines across all tasks, though to a lesser extent, indicating that preference-based optimization incurs a non-trivial cost to general capabilities. MAPS maintains performance close to the baseline across all benchmarks, demonstrating that it leaves the model's reasoning and general visual capabilities largely intact. Unlike prior suppression methods that sacrifice general capabilities for prior resistance, MAPS acquires more comprehensive world knowledge through rare-but-real supervision, preserving the model's understanding of the real world.

This preservation effect likewise generalizes across models and architectures. Extending MAPS to six additional models leaves performance on all four generalization benchmarks within a narrow margin, with the largest absolute change reaching only 0.87 points and no consistent direction of degradation, as reported in Table 9 of Appendix C.

Ablation on Prior Probing Threshold. To isolate the effect of the threshold α from training data quantity, we control the total training data to match that of $\alpha = 0.90$. For lower thresholds where more priors are selected, we proportionally reduce the number of rare-but-real descriptions per prior; for higher thresholds where fewer priors are selected, we prompt Claude Opus 4.7 to generate additional descriptions per prior to reach the same total. Table 5 reports results across four threshold values. A low threshold such as $\alpha = 0.50$ selects a large number of priors indiscriminately, including many that the model does not strongly hold, which interferes with targeted prior softening and leads to suboptimal performance on both ViSiP-Bench and generalization benchmarks. A high threshold such as $\alpha = 0.95$ is overly restrictive, covering too few priors to achieve meaningful improvement in visual perception. $\alpha = 0.90$ achieves the best

balance, maximally improving visual perception while preserving the model's reasoning and general capabilities.

6 Conclusion

We present ViSiP-Bench, a redesigned benchmark for evaluating language prior dependency in MLLMs. Existing benchmarks rely on a shared-prior assumption, yet cross-model prior agreement falls below 80%, and accuracy overestimates prior resistance as over 15% of items can be answered without any image, motivating a re-examination of how prior resistance should be measured.

ViSiP-Bench addresses these issues through a model-aware dynamic prior induction protocol paired with three complementary metrics. PAR and Filtered Accuracy capture prior resistance at the decision level by conditioning on each model's own elicited prior, while VIG provides a continuous measure of the confidence shift induced by visual evidence. Evaluation of 24 models reveals that prior adherence is pervasive across model families, does not diminish with scale, and manifests as VIG Polarization in larger models, which either override their priors decisively or adhere to them just as firmly. These findings suggest that inference-time interventions alone are insufficient, motivating a training-based solution that avoids indiscriminately suppressing the model's world knowledge.

We therefore propose MAPS, which identifies each model's most strongly held priors and softens them by enriching the model with targeted, factually grounded knowledge of rare-but-real scenarios, rather than suppressing priors indiscriminately. MAPS achieves substantial improvements in PAR and VIG on ViSiP-Bench while preserving performance on reasoning and general vision benchmarks, and this behavior holds consistently across seven models spanning three architecture families, demonstrating that enhanced visual reliance can be obtained without sacrificing a model's broader capabilities.

Limitations

First, the six categories in ViSiP-Bench, though designed to broadly cover commonsense knowledge relevant to MLLMs, follow a fixed taxonomy and cannot capture every form of world knowledge a model may rely on. Priors rooted in specialized domains or in rapidly evolving knowledge fall outside this taxonomy, and whether the patterns of prior adherence we observe extend to such cases remains an open question for future benchmarks to address.

Second, VIG relies on token-level output probabilities and is therefore inapplicable to closed models accessed only through APIs. This leaves the VIG Polarization phenomenon reported in this work unverified for proprietary models, which may exhibit different confidence dynamics than what we observe here.

Ethical Considerations

Data construction relies on proprietary generative models, including Claude Opus 4.7 for prior enumeration and scenario generation, and GPT-Image-2 and Nano Banana 2 for image synthesis, all accessed through official APIs in accordance with the respective providers' usage policies. Evaluation additionally involves querying proprietary models such as GPT-5.5, Claude Sonnet 4.6, and Gemini-3-flash-preview through their official APIs, again in compliance with the applicable terms of service. Content synthesized by generative models, as well as web-sourced images, is subject to the same three-criteria quality review, which additionally serves to filter out any inappropriate or misleading generations before inclusion in ViSiP-Bench. Human annotation involved in question and option design, as well as quality verification, was carried out by voluntary annotators who were fully informed of the task and compensated fairly for their time; the annotation task does not involve exposure to violent, explicit, or otherwise harmful content, as all images depict ordinary, if uncommon, real-world scenarios.

We note that generative AI tools were employed only for editing purposes, such as improving wording and sentence clarity, and played no role in the underlying research. All ideas, experiments, analyses, and conclusions presented in this paper originate from and were validated by the authors.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.U24A20250), Sichuan Science and Technology Program (No.2025ZNSFSC1487), the Fundamental Research Funds for the Central Universities (No.ZYGX2024J022 and No.ZYGX2024Z005), and the Science, Technology and Innovation Project of Shenzhen Longhua District (No.20260309G23410662).

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Anthropic. 2026a. Claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6. Accessed: 2026-05-26.

Anthropic. 2026b. Introducing claude opus 4.7. https://www.anthropic.com/news/claude-opus-4-7. Accessed: 2026-05-26.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Michal Golovanevsky, William Rudman, Michael A Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. 2025. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24848–24863.

Google. 2025. Gemini 3 flash preview. https://blog.google/products/gemini/gemini-3-flash/. Accessed: 2026-05-26.

Google. 2026. Build with Nano Banana 2: our best image generation and editing model. https://blog.google/innovation-and-ai/technology/developers-tools/build-with-nano-banana-2/. Accessed: 2026-05-26.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

Junlin Han, Shengbang Tong, David Fan, Yufan Ren, Koustuv Sinha, Philip Torr, and Filippos Kokkinos. 2026. Learning to see before seeing: Demystifying llm visual priors from language pre-training. In *International Conference on Learning Representations*, volume 2026, pages 61025–61055.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Filip Ilievski, Alessandro Oltramari, Kaixin Ma, Bin Zhang, Deborah L. McGuinness, and Pedro Szekely. 2021. Dimensions of commonsense knowledge. *Knowledge-Based Systems*, 229:107347.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. 2025. VLind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4129–4144, Albuquerque, New Mexico. Association for Computational Linguistics.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

Lin Long, Changdae Oh, Seongheon Park, and Sharon Li. 2026. Understanding language prior of lvlms by contrasting chain-of-embedding. In *International Conference on Learning Representations*, volume 2026, pages 94723–94754.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. 2024. Probing visual language priors in vlms. *arXiv preprint arXiv:2501.00569*.

OpenAI. 2026a. Gpt-5.5. https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-05-26.

OpenAI. 2026b. Introducing ChatGPT Images 2.0. https://openai.com/blog/introducing-chatgpt-images-2-0. Accessed: 2026-05-26.

Qwen Team. 2025a. Qwen2.5-vl. https://qwen.ai/blog?id=qwen2.5-vl.

Qwen Team. 2025b. Qwen3vl. https://qwen.ai/blog?id=qwen3-vl.

Qwen Team. 2026. Qwen3.5. https://qwen.ai/blog?id=qwen3.5.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Yixuan Wu, Yang Zhang, Jian Wu, Philip Torr, and Jindong Gu. 2026. Postalign: Multimodal grounding as a corrective lens for mllms. In *International Conference on Learning Representations*, volume 2026, pages 78875–78894.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. 2026a. Visulogic: A benchmark for evaluating visual reasoning in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pages 25966–26003.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. 2026b. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *Advances in Neural Information Processing Systems*, 38:82878–82905.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.

YiFan Zhang, Yang Shi, Weichen Yu, Qingsong Wen, Xue Wang, Wenjing Yang, Zhang Zhang, Liang Wang, and Rong Jin. 2025a. Debiasing multimodal large language models via penalization of language priors. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4232–4241.

Zefeng Zhang, Hengzhu Tang, Jiawei Sheng, Zhenyu Zhang, Yiming Ren, Zhenyang Li, Dawei Yin, Duohe Ma, and Tingwen Liu. 2025b. Debiasing multimodal large language models via noise-aware preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9423–9433.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2025. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *International Conference on Learning Representations*, volume 2025, pages 54415–54439.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Data Quality Analysis

A.1 Identification of No-Image Correct Responses

Over 15% of items in existing benchmarks can be answered correctly without any image. To identify the source of each such response and categorize it into one of three types (*Chance Success*, *Low-quality Framing*, or *Anomalous Prior*), we adopt a hybrid pipeline combining automated screening with human verification.

The pipeline proceeds as follows. All items answered correctly under the Blank condition are first collected as candidates. Each candidate then undergoes the following three-stage procedure.

Stage 1: Human Verification of Question Quality. Human annotators examine each candidate question along two dimensions. First, they assess whether the question is associated with a sufficiently strong and unambiguous commonsense prior. Questions where the correct answer is genuinely uncertain even without an image (for example, "what color is the dragon fruit?" where both red and white are valid) are flagged as *Low-quality Framing* due to an insufficiently strong prior. Second, annotators check whether the question wording contains implicit hints that guide the model toward the anomalous answer. For instance, a question such as "what is unusual about the color of the rose in this image?" implicitly signals that the rose's color is atypical, nudging the model toward selecting an uncommon option such as black or white. Such items are also categorized as *Low-quality Framing*.

Stage 2: Automated Probability Analysis. Items that pass Stage 1 (i.e., those with a strong and unambiguous prior and no implicit hints) are further analyzed by re-querying the target model under the Blank condition and extracting the full probability distribution over all answer options via softmax over the output logits. If the variance across option probabilities is low, indicating that the model assigns similar probabilities to all options, the item is categorized as *Chance Success*. This pattern is particularly prevalent in smaller models with weaker prior knowledge, where correct responses

Source	Model	Question	Reason for Categorization
Chance Success	Qwen3-VL-2B	When is the vernal equinox? (A. March 20 B. June 20 C. July 23 D. December 21)	Model assigns near-uniform probabilities: A: 28.3%, B: 24.1%, C: 26.8%, D: 20.8%, indicating no strong prior and random correct selection.
Low-quality Framing	—	What is unusual about the color of the rose in this image? (A. Red B. Black C. Pink D. White)	The phrase "unusual" implicitly hints at an atypical answer, guiding the model toward anomalous options.
Low-quality Framing	—	What color is the watermelon? (A. Red B. Green C. White D. Yellow)	The question does not specify flesh or rind, making the prior ambiguous and the correct answer uncertain without the image.
Anomalous Prior	Qwen2.5-VL-3B	What color is the dolphin? (A. Blue B. White C. Pink D. Grey)	Model assigns high probability to "Pink", indicating it inherently holds the anomalous answer as its own prior.

Table 6: Representative examples of the three sources of no-image correct responses.

arise from random selection rather than genuine prior activation. In our implementation, if the maximum probability among all options falls below 65%, the item is categorized as *Chance Success*.

Stage 3: Anomalous Prior Identification. If the probability distribution exhibits high variance, with one option receiving substantially higher probability than the others, the model is judged to hold a strong prior for this question. Since the model answers correctly under the Blank condition despite the item being designed to conflict with commonsense expectations, this indicates that the model inherently treats the anomalous answer as its own prior. Such items are categorized as *Anomalous Prior*.

A.2 Typical Cases

Table 6 provides representative examples of each source type identified by the above pipeline.

B Visual Style Examples

ViSiP-Bench incorporates images across four distinct visual styles to ensure that model performance is not confounded by image style or source. Table 7 lists the generation prompts used for each style, and Figure 7 presents one representative example from each style.

Style	Generation Prompt
Photorealistic	A police officer in full uniform stands at the front of a classroom, sharing career experiences with attentive students seated at desks.
Anime	A color-blind boy proudly holds up his drawing of a figure with green skin, smiling with satisfaction, in a vibrant anime art style.
Oil painting	A young boy stares in disbelief at a perfectly square watermelon sitting on a wooden table, his mouth wide open in shock, rendered in the style of a classical oil painting.
Sci-fi	A boy uses his smartphone to crack open a walnut, pressing the screen edge against the shell with focused determination, in a futuristic sci-fi setting with neon lighting.

Table 7: Generation prompts for each visual style in ViSiP-Bench.

C Cross-Model and Cross-Family Validation of MAPS

To verify that the effectiveness of MAPS is not specific to a single backbone, we apply MAPS to six additional models spanning three architecture families: InternVL3.5 (2B, 8B), InternVL3 (14B), Qwen2.5-VL (7B), Qwen3.5 (9B), and LLaVA-OneVision (7B). All models are fine-tuned under an identical configuration (prior probing threshold



(a) Photorealistic: A police officer shares career experiences in a classroom.



(b) Anime: A color-blind boy proudly displays his drawing of a green-skinned figure.



(c) Oil painting: A boy stares in disbelief at a square watermelon.



(d) Sci-fi: A boy cracks a walnut using his smartphone.

Figure 7: Representative examples of the four visual styles in ViSiP-Bench. Each image depicts a rare-but-real scenario designed to conflict with commonsense priors.

$\alpha = 0.90$, LoRA fine-tuning for 2 epochs), with no per-model hyperparameter adjustment, to verify that the improvement generalizes without model-specific tuning.

Table 8 reports results on ViSiP-Bench. Across all six models, MAPS consistently reduces PAR (by 3.44 to 8.06 points) and improves Filtered Accuracy (by 3.62 to 8.15 points), with no exceptions across model scale (2B–14B) or family. FC increases only marginally (by 2 to 7 samples), in sharp contrast to the collapse observed for Post-hoc in Table 3 (147 $\rightarrow$ 385), confirming that MAPS softens rather than indiscriminately erodes prior knowledge. VIG improves substantially for every model, in most cases more than doubling, indicating that the effect is not confined to a single backbone but reflects a general property of the model-aware prior softening procedure. We note that the absolute PAR reduction for LLaVA-OneVision-7B (-4.09) is somewhat smaller than for other models; this model also exhibits the highest base PAR (67.10) among the six, suggesting that models with a very large number of strongly held priors may require a longer fine-tuning schedule or a larger rare-but-real corpus to achieve proportionally similar gains.

Table 9 reports generalization results on the same six models. The largest absolute change across all 24 model–benchmark pairs is 0.87 points (MMStar, InternVL3.5-2B), and changes are not consistently signed: some models show small gains on certain benchmarks and small losses on others, with no systematic degradation pattern. This indicates that the preservation of general reasoning and visual understanding capabilities observed for Qwen3-VL-8B-Instruct in Table 4 is not an artifact of that particular model, but a consistent property of MAPS across scales and architecture families.

D Annotation Guidelines

Task Description. You will be presented with an image depicting a rare-but-real scenario that conflicts with commonsense expectations. Your task is to design a multiple-choice question that requires the respondent to carefully observe the image in order to answer correctly. The correct answer must be visually grounded in the image, and must differ clearly from what one would expect based on commonsense knowledge alone.

What is a Commonsense Prior? A commonsense

Model	Type	PAR↓	Acc↑	FC	$\mathbf{VIG_{avg}}$↑	$\mathbf{VIG_{med}}$↑
InternVL3.5-2B	base	51.51	42.94	182	4.41	4.11
	+MAPS	**46.91**	**47.11**	189	**8.91**	**9.20**
InternVL3.5-8B	base	47.17	44.64	156	7.13	7.48
	+MAPS	**39.11**	**51.85**	158	**15.94**	**15.63**
InternVL3-14B	base	48.40	41.33	161	6.21	6.33
	+MAPS	**42.60**	**47.91**	166	**13.81**	**13.48**
LLaVA-OV-7B	base	67.10	25.36	317	3.39	2.48
	+MAPS	**63.01**	**29.81**	319	**5.94**	**6.11**
Qwen2.5-VL-7B	base	55.35	34.51	198	5.56	4.93
	+MAPS	**51.91**	**38.13**	201	**10.63**	**11.21**
Qwen3.5-9B	base	51.24	39.36	137	5.24	5.24
	+MAPS	**43.22**	**47.51**	141	**12.97**	**13.62**

Table 8: Cross-model results on ViSiP-Bench. MAPS consistently reduces PAR and increases Filtered Accuracy and VIG across all six models spanning three architecture families, while FC remains close to the base model, indicating that prior knowledge is not indiscriminately disrupted.

Model	Type	MMMU-Pro	VisuLogic	MMStar	RealWorldQA
InternVL3.5-2B	base	34.10	24.10	61.18	63.12
	+MAPS	33.91	24.15	62.05	62.49
InternVL3.5-8B	base	41.20	28.60	65.23	67.32
	+MAPS	41.13	28.71	64.79	67.51
InternVL3-14B	base	40.10	26.30	64.14	63.98
	+MAPS	40.23	26.41	64.87	63.48
LLaVA-OV-7B	base	24.10	25.30	56.13	55.30
	+MAPS	24.18	26.11	55.84	55.15
Qwen2.5-VL-7B	base	38.30	26.00	59.14	61.38
	+MAPS	38.19	25.63	59.71	60.91
Qwen3.5-9B	base	42.10	28.61	66.51	68.31
	+MAPS	42.34	27.90	66.90	68.44

Table 9: Generalization results on reasoning (MMMU-Pro, VisuLogic) and general vision (MMStar, RealWorldQA) benchmarks across six models. MAPS preserves performance within a narrow margin for every model, with no consistent direction of degradation.

prior is a default expectation about the world that a person (or model) holds based on general knowledge. For example, most people expect swans to be white. If an image shows a black swan, a respondent relying on their prior rather than the image would incorrectly answer "white". Questions should be designed so that relying on the prior leads to a wrong answer, while carefully observing the image leads to the correct one.

Question Design Requirements.

(1) Identify the anomalous element in the image and formulate a question that directly targets it.

(2) Phrase the question in a neutral and objective manner. Do **not** use any wording that hints at the anomalous nature of the image (e.g., avoid phrases such as "what is unusual about...").

(3) Provide exactly four answer choices. One choice must correspond to the visually correct answer; the remaining three must be plausible distractors grounded in commonsense expectations.

(4) Ensure that no option is obviously implausible, as this would allow the correct answer to be identified by elimination.

Quality Review Criteria. Each sample must satisfy all three of the following criteria:

(1) The image is visually clear and unambiguously supports a single correct answer.

(2) All answer choices are directly relevant to the question and cannot be eliminated by common sense alone.

(3) The question wording contains no hint of anomaly that would suppress prior activation before the respondent views the image.

Common Mistakes. Please avoid the following common errors:

(1) *Implicit anomaly hints*: Phrasing such as "what is unusual about the color of the rose?" signals that something is atypical, guiding the respon-

Model	Physical	Time & Space	Scientific	Customs & Culture	Psychology	Function
Open-source Models						
InternVL3.5-2B	34.1	49.2	69.4	51.3	43.2	60.1
InternVL3.5-8B	30.2	44.1	63.2	47.1	39.4	55.2
InternVL3.5-14B	22.1	29.8	49.7	29.1	21.3	39.4
InternVL3.5-38B	16.3	21.4	38.9	21.4	19.7	21.1
InternVL3-2B	40.2	55.3	78.4	58.2	51.3	68.2
InternVL3-14B	28.4	41.2	62.1	44.3	37.2	52.1
InternVL3-38B	20.1	31.4	52.3	33.2	28.4	41.2
InternVL3-78B	18.2	29.1	49.8	31.4	26.1	38.9
Qwen2.5-VL-3B	41.3	56.2	79.1	59.4	52.1	69.3
Qwen2.5-VL-7B	33.2	48.1	68.3	50.2	42.3	59.1
Qwen2.5-VL-32B	24.3	36.2	58.4	38.1	31.2	47.3
Qwen2.5-VL-72B	19.4	30.1	51.2	32.3	27.4	40.1
Qwen3-VL-2B	38.2	52.1	74.3	55.1	48.2	64.3
Qwen3-VL-4B	29.3	43.2	65.1	45.2	38.3	54.2
Qwen3-VL-8B	21.4	32.1	54.2	34.3	29.1	43.4
Qwen3.5-2B	34.3	48.2	70.1	51.2	44.3	61.2
Qwen3.5-9B	28.2	42.1	64.3	44.1	37.2	53.3
Qwen3.5-27B	20.3	30.2	52.1	32.4	27.3	41.3
LLaVA-OV-0.5B	48.2	63.1	84.3	65.2	58.1	74.2
LLaVA-OV-7B	43.1	58.2	79.4	60.3	53.2	69.1
LLaVA-OV-72B	36.2	51.3	73.2	53.4	46.3	62.4
Proprietary Models						
GPT-5.5	12.1	11.9	35.2	15.4	14.2	25.1
Claude Sonnet 4.6	13.4	19.4	36.1	18.3	19.1	24.3
Gemini-3-flash	12.9	18.1	38.2	14.2	13.4	27.1

Table 10: Per-category PAR results for all evaluated models. Scientific & Mathematic consistently exhibits the highest PAR values, while Physical Properties proves most amenable to visual override.

dent toward the anomalous answer without looking at the image.

(2) *Eliminable distractors*: Including options that are obviously implausible allows the correct answer to be identified by elimination rather than visual observation.

(3) *Ambiguous reference*: Questions such as "what color is this?" without specifying the target object introduce ambiguity. Always specify the object clearly (e.g., "what color is the watermelon flesh?").

Disagreement Resolution. If the two reviewers disagree on whether a sample passes review, the case is escalated to a third arbitrator whose decision is final. All samples must reach unanimous agreement before inclusion in the dataset.

E Per-Category Analysis

Table 10 presents the per-category PAR results for all evaluated models. *Scientific & Mathematic* consistently exhibits the highest PAR values across model families, while *Physical Properties* proves most amenable to visual override. This pattern remains largely consistent across models, suggesting that category-level difficulty is an intrinsic property of the prior type rather than a model-specific characteristic.

F Prompt Template

To ensure a fair comparison between the Blank and Visual conditions, we use an identical prompt template for both. The prompt is designed for image-based questions, with the instruction to guess based on prior knowledge when no image is visible. This design ensures that any difference in model responses between the two conditions is attributable solely to the presence or absence of the image, rather than any difference in the prompt.

The prompt template used in both conditions is as follows:

```
In this picture, [QUESTION]
A. [OPTION A]
B. [OPTION B]
C. [OPTION C]
D. [OPTION D]
If you cannot see the image, please make your best guess among all answer choices based on your own judgment.
```