POLARVLM: WHEN POLARIZATION MEETS VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mainstream vision-language models (VLMs) fundamentally struggle with severe optical ambiguities, such as reflections and transparent objects, due to the inherent limitations of standard RGB inputs. While polarization imaging captures polarimetric physical parameters that resolve these ambiguities, existing methods are constrained by fixed-format outputs and remain isolated from open-ended reasoning. To bridge this semantic-physical gap, we introduce **PolarVLM**, the first multimodal framework integrating polarimetric physical parameters into VLMs. By employing a dual-stream architecture and a progressive two-stage training strategy, PolarVLM effectively prevents physical misinterpretations while largely retaining general visual abilities. Complementing our architecture, we construct **PolarVQA**, the first benchmark for polarization-aware VQA, featuring 75K physics-grounded instruction-tuning pairs targeting reflective and transparent scenes. Experiments show that PolarVLM surpasses the RGB baseline by 25.4% overall across five evaluation tasks, with remarkable gains of 26.6% in reflection recognition and 34.0% in glass counting, successfully unlocking physics-aware semantic understanding.

1 INTRODUCTION

Mainstream vision-language models (VLMs) excel at open-ended visual question answering (VQA) using RGB inputs (Liu et al., 2023; Li et al., 2023; Liu et al., 2024a). However, real-world environments frequently feature specular reflections and transparent objects, which pose systematic challenges to this paradigm. These phenomena introduce inherent *optical ambiguities*: because standard RGB sensors capture only intensity and color, they cannot physically disentangle actual objects from reflected virtual content, nor perceive transparent surfaces lacking distinct visual boundaries. Relying exclusively on such ambiguous photometric features, existing VLMs are highly susceptible to severe perceptual illusions. Consequently, as illustrated by the RGB-only failures in Fig. 1, even state-of-the-art models fail to ground accurate semantic reasoning in these common physical scenarios.

Polarization imaging breaks this bottleneck by capturing polarimetric physical parameters (*e.g.*, Degree of Linear Polarization (DoLP) and Angle of Linear Polarization (AoLP)) that remain invisible to standard RGB sensors (Goldstein, 2017; Collett, 2005). While these parameters have proven effective for low-level vision tasks like reflection removal (Lyu et al., 2019; Wang et al., 2025b; Yao et al., 2025) and transparent object segmentation (Mei et al., 2022; Shao et al., 2023), their potential for high-level semantic reasoning remains largely untapped. This is because existing approaches are strictly constrained by fixed-format outputs, fundamentally preventing them from supporting open-ended dialogue. Moreover, a critical void of polarization-text instruction data leaves modern VLMs entirely blind to these physical parameters. Therefore, there is a compelling need to develop a unified multimodal framework and benchmark that jointly integrate RGB and polarimetric cues, injecting physics-aware perception into open-ended VQA to unlock accurate semantic understanding.

Unlocking this capability, however, presents non-trivial challenges spanning both model architecture and data alignment. On the architectural front, directly injecting polarimetric physical parameters into standard VLMs leads to a severe representation gap. The physical properties encoded by these parameters differ from the appearance-based semantic features expected by pretrained visual encoders. A naive workaround is a cascaded approach that utilizes polarization solely for low-level image preprocessing. However, this strategy creates an opaque pipeline where early estimation errors

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

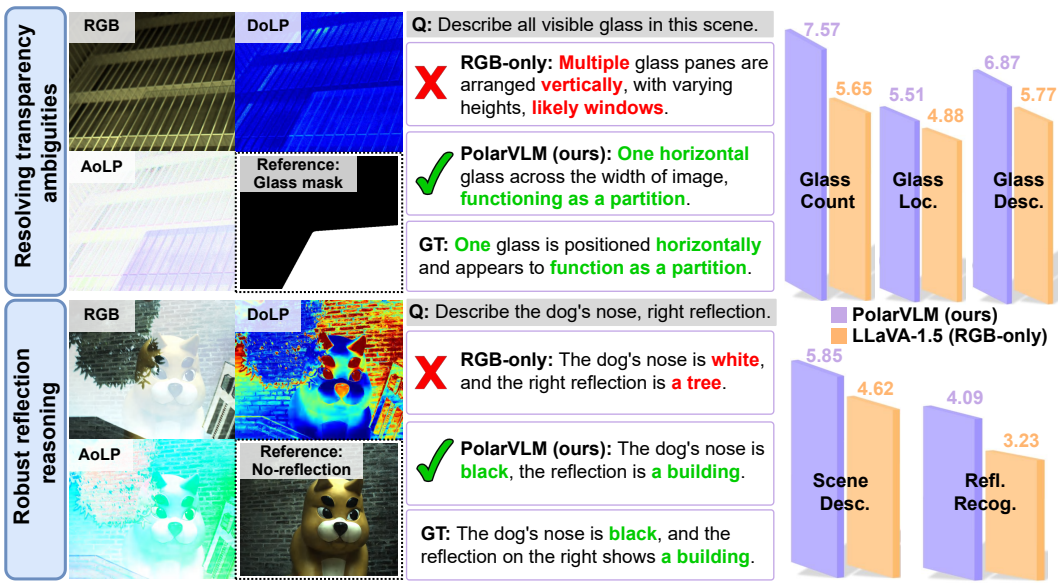

Figure 1: PolarVLM overcomes optical ambiguities through physics-aware multimodal reasoning. By integrating polarimetric physical parameters, PolarVLM significantly outperforms conventional RGB-only baselines (*e.g.*, fine-tuned LLaVA-1.5 (Liu et al., 2024a)) in resolving transparency ambiguities and performing robust reflection reasoning, both qualitatively and quantitatively.

inevitably accumulate. Consequently, the language model remains isolated from the raw physical signals, precluding genuine physics-aware reasoning. It is therefore critical to design a dedicated fusion mechanism that preserves the distinct strengths of both modalities. Concurrently, on the data front, existing polarimetric datasets offer only pixel-level annotations without language descriptions. The challenge lies in bridging this semantic-physical gap by translating dense physical measurements into high-fidelity textual supervision. Crucially, these annotations must strictly demonstrate the boundary of what polarization can actually express (*e.g.*, geometric structure and material properties). This precision is essential to prevent the model from misinterpreting physical parameters as ordinary visual textures, or erroneously associating them with unrelated semantic concepts.

To address these limitations, we introduce **PolarVLM**, the first multimodal framework specifically designed to bridge the semantic-physical gap in open-ended VQA. By anchoring the language model in polarimetric physics, PolarVLM effectively resolves severe optical ambiguities for reflections and transparent objects (see Fig. 1). Overcoming the isolation of cascaded approaches, we construct a dual-stream architecture upon LLaVA-1.5 (Liu et al., 2024a). This design injects physical measurements via a dedicated polarization branch while robustly preserving pretrained RGB semantic priors. The resulting heterogeneous features are integrated via token-level concatenation, empowering the language model's self-attention layers to dynamically cross-reference physics-aware and appearance-based tokens. To prevent the model from misinterpreting physical parameters as ordinary visual textures, we propose a progressive two-stage training strategy. By initially aligning polarimetric features to the semantic space prior to dual-stream joint instruction tuning, we ensure that physical signals correctly ground the model's reasoning without disrupting its inherent visual-linguistic capabilities.

Complementing our architectural design, we introduce **PolarVQA** to bridge the critical void in polarimetric instruction data. We formulate a physics-guided data generation pipeline that integrates visual foundation models (Xiao et al., 2024), exact polarimetric computations, and large language models (Hurst et al., 2024). This pipeline successfully translates low-level physical measurements into high-fidelity textual supervision while strictly preventing the introduction of unrelated semantic concepts. Built upon raw multimodal captures from public datasets (Yao et al., 2025; Mei et al., 2022), we construct 28.5K physics-grounded captions and 46.8K diverse instruction-tuning pairs. Structurally, the benchmark is unified under two core dimensions: *resolving transparency ambiguities* and *robust reflection reasoning*, which feature a dedicated *counterfactual reasoning* subset explicitly

designed to penalize reflection-induced object hallucinations. By addressing these severe optical ambiguities, PolarVQA establishes the first comprehensive benchmark for training and evaluating physics-aware vision-language models.

In summary, our primary contributions include:

- **PolarVLM**, the first multimodal framework integrating polarimetric physical parameters into VLMs for open-ended VQA under severe optical ambiguities.
- **A dual-stream architecture and progressive training strategy** to bridge the semantic-physical gap, preventing physical misinterpretations while largely retaining general visual abilities.
- **PolarVQA**, the first benchmark for polarization-aware VQA, providing 75K physics-grounded captions and instruction-tuning pairs targeting reflective and transparent scenes.

2 Related work

Polarization-based vision. Since the polarization properties of light are fundamentally governed by surface normals, textures, and material characteristics, polarization has become an indispensable physical cue for overcoming the limitations of conventional appearance-based approaches. In the realm of 3D vision, polarization has been widely exploited to solve a variety of ill-posed problems, including shape and normal estimation (Deschaintre et al., 2021; Ba et al., 2020; Lyu et al., 2023; Lei et al., 2022; Cao et al., 2023), 3D reconstruction and inverse rendering (Han et al., 2024; Chen et al., 2024a; Dave et al., 2022), and depth sensing (Kadambi et al., 2017; Tian et al., 2023). Furthermore, because reflection and scattering consistently alter the polarization state of light, polarimetric cues have been extensively applied to low-level vision tasks such as reflection removal (Lyu et al., 2019; Wang et al., 2025b; Yao et al., 2025; Lyu et al., 2022; Lei et al., 2020), dehazing (Schechner et al., 2001; Zhou et al., 2021), high dynamic range imaging (Wu et al., 2020; Zhou et al., 2023), color constancy (Ono et al., 2022), and shadow removal (Zhou et al., 2025b). Most critically for complex real-world scenes, polarization provides unique discriminative features for transparent object perception (Mei et al., 2022; Shao et al., 2023; Kalra et al., 2020), scenarios where purely RGB-based methods suffer from severe generalization bottlenecks due to the absence of explicit physical constraints (Zhang et al., 2018; Wan et al., 2019). Despite their significant progress in these respective domains, existing methods remain confined to task-specific modules with fixed-format outputs, while PolarVLM pioneers the translation of polarimetric physical quantities into language supervision to enable open-ended, physics-grounded visual question answering.

Vision-language models. Recent VLMs have successfully aligned pretrained visual encoders with large language models. Pioneer frameworks like LLaVA (Liu et al., 2023; 2024a) and MiniGPT-4 (Zhu et al., 2023) project visual tokens via multi-layer perceptrons, while BLIP-2 (Li et al., 2023) and InstructBLIP (Dai et al., 2023) employ learnable queries and cross-attention mechanisms. More recently, Qwen2-VL (Wang et al., 2024) and InternVL (Chen et al., 2024c) have introduced dynamic resolution and scaled language backbones. However, these general-purpose VLMs rely merely on RGB inputs. Their perceptual failures in reflective and transparent scenarios stem not from insufficient model capacity, but from the lack of physical information within the RGB modality itself.

Incorporating non-RGB modalities. To transcend the physical limitations of RGB, researchers have increasingly explored extending vision-language alignment to diverse modalities. Foundational efforts successfully extended Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) to point clouds (Zhang et al., 2022), video (Xu et al., 2021), and audio (Guzhov et al., 2022), while unified frameworks like ImageBind (Girdhar et al., 2023) have established shared embedding spaces across multiple sensory inputs. Furthermore, domain-specific sensors are progressively being integrated into VLMs to tackle extreme conditions; notable examples include tactile sensing for physical deformation (Yang et al., 2024), event cameras for high-dynamic scenes (Li et al., 2025), and thermal imaging for low-light environments (Jiang et al., 2024). Despite successes with other modalities, incorporating polarization into VLMs is hindered by its complex information overlap with RGB and a critical lack of text-aligned datasets. PolarVLM resolves these challenges, introducing the first dual-stream framework and companion PolarVQA benchmark for physics-grounded open-ended reasoning.

3 THE POLARVLM FRAMEWORK

3.1 PHYSICAL IMAGE FORMATION AND POLARIMETRIC ENCODING

To introduce physical perception into vision-language models, it is essential to first bridge the gap between raw optical measurements and neural representations. Standard polarization cameras employ a micro-polarizer array to capture the modulation of light. According to Malus's law (Collett, 2005), the measured intensity I_α passing through a linear polarizer oriented at angle α can be expressed in terms of the Stokes parameters $\mathbf{S} = [S_0, S_1, S_2]^\top$:

$$I_\alpha = \frac{1}{2}S_0 + \frac{1}{2}S_1\cos(2\alpha) + \frac{1}{2}S_2\sin(2\alpha), \tag{1}$$

where S_0 denotes the total light intensity, and S_1 and S_2 characterize the linearly polarized components. In a single exposure, the sensor captures four spatially aligned directional intensities: I_{0°, I_{45°, I_{90°, and I_{135°. Based on Eq. 1, the Stokes parameters can be strictly decoupled as:

$$S_0 = I_{0^\circ} + I_{90^\circ}, \quad S_1 = I_{0^\circ} - I_{90^\circ}, \quad S_2 = I_{45^\circ} - I_{135^\circ}. \tag{2}$$

Once the Stokes vector is obtained, the fundamental physical properties of the incident light, namely, the Degree of Linear Polarization (DoLP, denoted as P) and the Angle of Linear Polarization (AoLP, denoted as Φ), can be analytically derived:

$$P = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}, \quad \Phi = \frac{1}{2}\arctan\left(\frac{S_2}{S_1}\right). \tag{3}$$

Here, $P \in [0, 1]$ indicates the proportion of linearly polarized light, serving as a strong physical cue for separating reflections and revealing transparent surfaces. $\Phi \in [0, \pi)$ represents the oscillation orientation of the electric field. However, directly feeding Φ into a pretrained visual encoder introduces severe domain mismatch issues. Because the physical angle is π-periodic, values near 0 and π represent nearly identical physical states but exhibit maximal numerical distance. This angular discontinuity disrupts the continuous feature space expected by standard convolutional or transformer-based encoders.

To seamlessly inject these physical properties into the vision-language domain, we design a continuous, neural-friendly polarimetric representation. We map the discontinuous Φ into a continuous 2D Cartesian space via trigonometric transformations $\sin(2\Phi)$ and $\cos(2\Phi)$. Notably, this is physically equivalent to utilizing the normalized Stokes parameters $S_2/\sqrt{S_1^2 + S_2^2}$ and $S_1/\sqrt{S_1^2 + S_2^2}$. By concatenating these components with the DoLP, we construct a normalized polarization input:

$$\mathbf{X}_{\text{pol}} = [P,\ \sin(2\Phi),\ \cos(2\Phi)]. \tag{4}$$

This physics-grounded encoding strictly preserves the discriminative power of the original polarimetric measurements while eliminating periodic boundary ambiguities, explicitly aligning the physical signal distribution with the input requirements of the subsequent dual-stream architecture.

3.2 DUAL-STREAM ARCHITECTURE AND MODALITY FUSION

The core philosophy behind PolarVLM is to inject polarimetric physical perception into the language model without compromising its foundational semantic reasoning capabilities. Directly modifying the visual encoder to accept additional input channels would inevitably disrupt the pretrained semantic priors. Therefore, as shown in Fig. 2, we design a structurally symmetric dual-stream architecture built upon the LLaVA-1.5 (Liu et al., 2024a) framework, explicitly decoupling appearance processing from physical feature extraction. The RGB encoder and its projection layer remain completely frozen throughout training, providing an anchor for stable semantic understanding.

Physics-aware visual branch. To extract representations from the polarimetric input $\mathbf{X}_{\text{pol}}$, we construct a parallel visual branch adopting the same CLIP ViT-L/14 architecture (Radford et al., 2021) as the RGB pathway. Rather than training this branch from scratch, we initialize it with pretrained RGB weights. This design choice is motivated by the strong morphological affinities between polarimetric parameters and low-level visual features: DoLP resembles an edge map (Zhou et al., 2025a;d), while the spatial gradients of the trigonometric AoLP channels closely mimic a

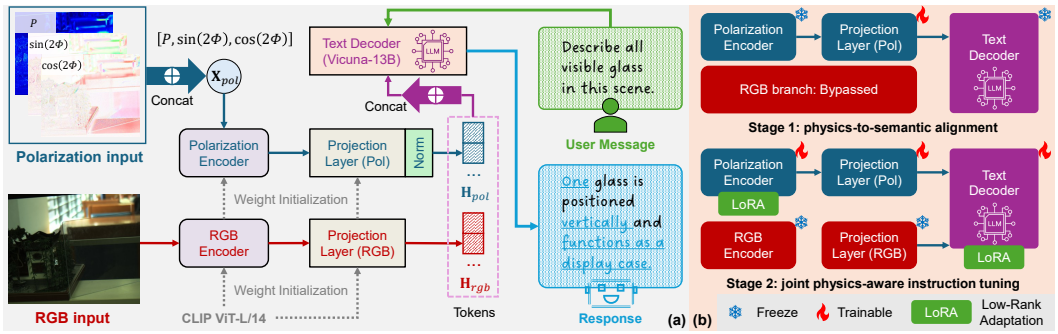

Figure 2: Overall design of PolarVLM. (a) Dual-stream physics-aware architecture with sequence-concatenation modality fusion. (b) Progressive two-stage training strategy to avoid modality collapse and achieve robust physical-semantic alignment.

surface normal map (Peters et al., 2023; Zhou et al., 2025c). Consequently, the edge detection and texture extraction priors inherently learned by the pretrained vision model can effectively transfer to the physical domain. The extracted physical features are then mapped into the LLM embedding space via an MLP projection layer. Since physical and appearance features differ in numerical distribution, we introduce a normalization layer at the polarization projector output to align latent magnitudes and prevent scale-induced attention imbalance.

LLM-delegated modality fusion. Given the two heterogeneous sets of visual tokens, $\mathbf{H}_{\text{rgb}}$ and $\mathbf{H}_{\text{pol}}$, an appropriate fusion mechanism is critical. Conventional fusion strategies exhibit distinct limitations in this multimodal context: residual addition implicitly assumes strict semantic additivity at corresponding spatial locations, which fundamentally violates the orthogonal physical meanings of RGB and polarization. Conversely, introducing auxiliary cross-attention modules yields excessive learnable parameters, risking severe overfitting when fine-tuning on limited multimodal data. To circumvent these issues, PolarVLM adopts sequence concatenation. By simply concatenating $\mathbf{H}_{\text{rgb}}$ and $\mathbf{H}_{\text{pol}}$ along the sequence dimension, we impose no restrictive prior assumptions. Instead, we elegantly delegate the cross-modal information routing entirely to the robust self-attention mechanism of the LLM. This allows the model to dynamically cross-reference physics-aware tokens specifically when the appearance-based tokens are corrupted by specular reflections or transparency.

4 Progressive Two-Stage Training Strategy

Training a heterogeneous multimodal framework from scratch is notoriously prone to modality collapse, where the language model relies entirely on familiar RGB representations and effectively ignores newly introduced physical signals. To avoid this and ensure robust physical grounding, PolarVLM employs a progressive two-stage training paradigm.

Stage 1: physics-to-semantic alignment. The primary objective of this stage is to strictly enforce the semantic grounding of the polarimetric signals. As illustrated in Fig. 2, we explicitly bypass the RGB branch and feed only the physics-aware tokens $\mathbf{H}_{\text{pol}}$ into the language model. This unimodal bottleneck design forces the polarization branch to establish an independent semantic representation, categorically preventing the model from developing a lazy reliance on RGB priors during early training. Architecturally, we freeze the LLM and the polarization encoder's transformer blocks, updating only the polarization patch embedding layer and the MLP projector. This design is highly deliberate: the patch embedding directly ingests the normalized physical values and thus requires domain adaptation to map continuous physical distributions into discrete semantic tokens, whereas the higher-level spatial reasoning capabilities of the ViT are domain-agnostic and should be preserved.

Stage 2: joint physics-aware instruction tuning. Having established a foundational semantic understanding of polarization, the model transitions to dual-stream instruction tuning, as shown in Fig. 2. Here, both modalities are activated, and the concatenated visual tokens $[\mathbf{H}_{\text{rgb}}, \mathbf{H}_{\text{pol}}]$ are jointly processed by the language model. The network is trained on complex, multi-turn VQA tasks involving reflections and transparency. To endow the model with advanced reasoning while

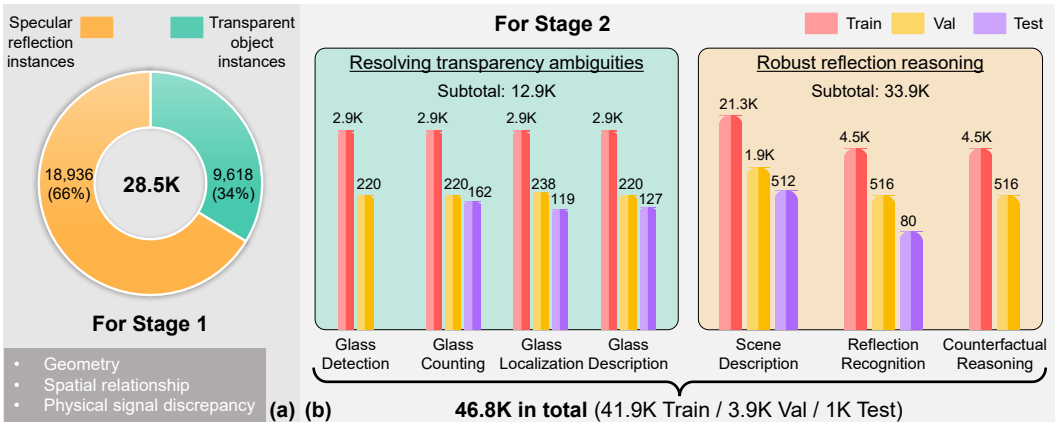

Figure 3: Statistical overview and task taxonomy of the PolarVQA benchmark. (a) For Stage 1, the distribution of high-fidelity polarization captions is partitioned by physical scenarios and explicitly annotated semantic variants. (b) For Stage 2, the diverse multi-turn instruction-tuning pairs are categorized across progressive cognitive tasks, with the y-axis presented in logarithmic scale to accommodate varying data magnitudes.

preventing catastrophic forgetting, we freeze the polarization patch embedding and instead adapt the polarization ViT's attention modules alongside the LLM via Low-Rank Adaptation (LoRA) (Hu et al., 2022). This creates a decoupled adaptation scheme: Stage 1 achieves input-level physical alignment, while Stage 2 realizes high-level cross-modal semantic adaptation.

Training objective. Throughout both stages, we optimize the model using the standard autoregressive next-token prediction objective. Let $\mathbf{H}$ denote the visual tokens provided to the language model, where $\mathbf{H} = \mathbf{H}_{\text{pol}}$ in Stage 1, and $\mathbf{H} = [\mathbf{H}_{\text{rgb}}, \mathbf{H}_{\text{pol}}]$ in Stage 2. Given the embedded instruction tokens $\mathbf{H}_{\text{instruct}}$, the model maximizes the likelihood of generating the target answer sequence:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{L} \log p_\theta(y_i \mid \mathbf{H}, \mathbf{H}_{\text{instruct}}, y_{<i}), \quad y = \{y_1, \ldots, y_L\}, \tag{5}$$

where θ denotes the trainable parameters of the respective stage, L is the sequence length, and $y_{<i}$ represents the previously generated tokens.

5 The PolarVQA Benchmark

To effectively train PolarVLM and rigorously evaluate its physics-aware reasoning capabilities, we construct PolarVQA, the first comprehensive vision-language benchmark targeting specular reflections and transparent objects. We source raw multi-angle polarimetric measurements and RGB images from two public datasets, PolarFree (Yao et al., 2025) and RGBP-Glass (Mei et al., 2022). However, since these datasets provide only pixel-level physical annotations without natural language descriptions, they cannot be directly utilized for vision-language instruction tuning.

Physics-guided data generation. Rather than relying on manual labeling, we design an automated, physics-guided data generation pipeline to translate raw physical measurements into high-fidelity textual supervision. At a high level, this pipeline adopts a visual foundation model (Florence-2-large (Xiao et al., 2024)) for spatial structural extraction, precise polarimetric computations for physical grounding, and a large language model (GPT-4o-mini (Hurst et al., 2024)) for final text synthesis. To prevent the language model from hallucinating and to avoid introducing spurious supervisory signals into the polarization modality, we explicitly strip RGB-specific semantics (*e.g.*, color and texture) from the polarization prompts. The detailed architecture of this generation pipeline is provided in Appendix C.

Dataset composition and task diversity. As illustrated in Fig. 3, the PolarVQA benchmark provides a rich, multi-tiered task distribution strictly partitioned into non-overlapping training, validation, and test splits. For Stage 1 (physics-to-semantic alignment), we generate 28.5K high-fidelity polarization

captions (25.2K train / 3.3K val). This split inherently balances optical challenges, comprising roughly 18.9K specular reflection instances and 9.6K transparent object instances, with each scene annotated across three semantic variants (geometry, spatial relationship, and physical signal discrepancy). For Stage 2 (joint physics-aware instruction tuning), we construct 46.8K diverse multi-turn instruction-tuning pairs (41.9K train / 3.9K val / 1K test) structured precisely around our two core evaluation dimensions. Under *resolving transparency ambiguities*, the evaluation encompasses four distinct tasks that progressively scale from fundamental perception to spatial reasoning: *glass detection*, *glass counting*, *glass localization*, and comprehensive *glass description*. Under *robust reflection reasoning*, the evaluation includes holistic *scene description* and *reflection recognition*, alongside a dedicated *counterfactual reasoning* subset. In this specific subset, we leverage spurious objects that exist solely in reflections to construct intentionally misleading instructions, explicitly training the model to resist object hallucinations by anchoring its reasoning in orthogonal polarization cues. This task design ensures that PolarVQA serves as a comprehensive testbed for multimodal physical perception.

6 EXPERIMENTS

Robustness and generalization. We validate PolarVLM's robustness through three independent LLM judges, a judge-free multiple-choice protocol, human evaluation, and multi-backbone generalization across four VLM architectures (Appendix B). We further confirm that polarimetric fine-tuning incurs only 3–6% degradation on MME (Fu et al., 2025), MMBench (Liu et al., 2024b), and MMStar (Chen et al., 2024b), with a modest gain on RealWorldQA (Appendix A).

6.1 EXPERIMENTAL SETUP

Implementation details. PolarVLM is initialized from the LLaVA-1.5-13B framework (Liu et al., 2024a), integrating Vicuna-13B as the language model backbone and CLIP ViT-L/14 as the visual encoder. To ensure computational efficiency, we optimize the model using QLoRA (Dettmers et al., 2023), where the frozen backbones are loaded in 4-bit NormalFloat (NF4) quantization and trainable parameters are updated under a BF16 mixed-precision regime. The training process strictly follows our proposed progressive schedule: Stage 1 aligns physical representations using 28.5K polarization captions, while Stage 2 performs joint instruction tuning on 41.9K PolarVQA multi-turn pairs. All experiments are conducted on a workstation equipped with two NVIDIA RTX 5880 Ada GPUs (48GB VRAM each). The comprehensive hyperparameter configuration is detailed in Appendix D.

Test benchmark and evaluation protocol. We evaluate all models on the PolarVQA test set, which comprises 1K multi-turn question-answer pairs across five distinct cognitive tasks. Structured around our two core dimensions, these encompass *glass counting*, *glass localization*, and *glass description* for resolving transparency ambiguities, alongside holistic *scene description* and *reflection recognition* for robust reflection reasoning. We adopt an LLM-as-a-Judge protocol (Wu et al., 2025; Jiang et al., 2026b), utilizing GPT-4o-mini (Hurst et al., 2024) to assign a quantitative score on a 1–10 scale based on semantic consistency, physical accuracy, and resistance to hallucinations; we report the sample-weighted average across tasks. Full protocol details are in Appendix E.

6.2 QUANTITATIVE EVALUATION

Compared baselines. To comprehensively evaluate the physical reasoning capabilities of PolarVLM, we benchmark against a diverse spectrum of multimodal pipelines (Tab. 1), categorized as follows:

- **Open-weight RGB-only VLMs:** Leading open-source frameworks, including BLIP2-Flan-T5-XL (Li et al., 2023), InternVL3-14B (Zhu et al., 2025) and Qwen2.5-VL-7B (Bai et al., 2025b).
- **Proprietary VLMs:** State-of-the-art closed-source models, specifically GPT-4.1 (Achiam et al., 2023).
- **Controlled baseline:** A strictly controlled LLaVA-1.5-13B (Liu et al., 2024a) fine-tuned on the exact same instruction data as our model, isolating the contribution of the polarization modality.
- **Cascaded polarimetric pipelines:** Two-stage approaches that apply task-specific physical preprocessing—such as reflection removal (PolarFree (Yao et al., 2025)) or transparent object masking (Glass Seg. (Mei et al., 2022))—prior to VLM inference.

Table 1: Quantitative evaluation on the PolarVQA test set. Scores are reported on a 1–10 scale using the LLM-as-a-Judge protocol. For cascaded pipelines, "–" denotes tasks strictly not supported by the pipeline's heuristics, and "N/A" indicates reflection cues are explicitly destroyed before evaluation.

Model / Pipeline	Resolving Transparency Ambiguities			Robust Reflection Reasoning		Overall
	Glass Count	Glass Loc.	Glass Desc.	Scene Desc.	Refl. Recog.	
Open-Weight RGB-only VLMs						
BLIP2-Flan-T5-XL (Li et al., 2023)	2.53	3.55	3.05	3.03	2.06	2.94
InternVL3-14B (Zhu et al., 2025)	4.55	5.40	6.55	5.02	2.81	5.01
Qwen2.5-VL-7B (Bai et al., 2025b)	5.22	5.38	6.65	5.54	3.10	5.42
Proprietary VLMs						
GPT-4.1 (Achiam et al., 2023)	4.81	5.46	5.96	5.66	3.71	5.38
Controlled Baseline (Trained on PolarVQA)						
LLaVA-1.5-13B (Liu et al., 2024a) (RGB-only)	5.65	4.88	5.77	4.62	3.23	4.85
Cascaded Polarimetric Pipelines						
PolarFree (Yao et al., 2025) + LLaVA-1.5-13B (Liu et al., 2024a)	–	–	–	4.34	N/A	–
PolarFree (Yao et al., 2025) + GPT-4.1 (Achiam et al., 2023)	–	–	–	5.71	N/A	–
Glass Seg. (Mei et al., 2022) + LLaVA-1.5-13B (Liu et al., 2024a)	2.94	4.29	4.54	–	–	–
Glass Seg. (Mei et al., 2022) + GPT-4.1 (Achiam et al., 2023)	7.16	5.48	6.54	–	–	–
End-to-End Multimodal Pipeline						
PolarVLM (Ours)	**7.57**	**5.51**	**6.87**	**5.85**	**4.09**	**6.08**

The necessity of polarimetric grounding. As reported in Tab. 1, PolarVLM significantly outperforms the directly comparable RGB-only baseline, elevating the Overall score from 4.85 to 6.08 (a relative gain of 25.4%). This consistent improvement across all physical reasoning tasks confirms that polarimetric cues provide critical structural evidence—such as DoLP responses induced by Fresnel reflections—that effectively resolves the visual ambiguities inherent in RGB imagery. Remarkably, PolarVLM also surpasses vastly larger general-purpose VLMs (including GPT-4.1) on physics-heavy tasks like *glass localization* and *reflection recognition*. This demonstrates that scaling model parameters cannot compensate for the lack of orthogonal physical sensing modalities. Paired bootstrap tests confirm that these gains are statistically significant across all tasks ($p < 0.01$; Appendix B.5).

End-to-end vs. cascaded pipelines. Tab. 1 further reveals the information bottlenecks of cascaded pipelines. PolarFree (Yao et al., 2025) reflection removal destroys the optical cues required for *reflection recognition* (N/A), and even on *scene description*—where removal should theoretically simplify perception—this rigid filtering discards contextual physical evidence, causing the pipeline to underperform PolarVLM even when powered by GPT-4.1 (5.71 vs. 5.85). Similarly, relying on intermediate segmentation masks (Mei et al., 2022) introduces error propagation that limits transparency-related tasks (*e.g.*, *glass counting*). In contrast, PolarVLM preserves raw physical evidence, enabling joint appearance-polarimetric reasoning without fragile heuristics.

6.3 Qualitative analysis

Fig. 4 compares PolarVLM with the RGB-only LLaVA-1.5-13B (Liu et al., 2024a) baseline and GPT-4.1 (Achiam et al., 2023). Additional examples are in Appendix F; Appendix H tests whether rendered polarization visualizations suffice for unaligned VLMs.

Resolving transparency ambiguities. In the left column, both baselines misidentify transparent geometry (*e.g.*, labeling a vertical partition as a "horizontal tabletop") and mislocalize glass positions. By grounding reasoning in DoLP and AoLP contours, PolarVLM accurately captures surface orientation and spatial layout.

Robust reflection reasoning. In the right column, the RGB-only model hallucinates spurious reflected objects, while GPT-4.1 misidentifies reflected structures. Under intense specular reflections (bottom right), baselines hallucinate incorrect colors (light blue, red and gold), whereas PolarVLM leverages the polarimetric discrepancy between reflected and Lambertian light to correctly describe the intrinsic appearance.

6.4 Ablation studies

Tab. 2 validates our key design choices. Removing Stage 1 degrades performance, confirming that grounding polarimetric features in the LLM's semantic space before instruction tuning is essential. Removing the normalization layer causes the most severe drop, as unbalanced activation scales between modalities fatally bias the LLM. Disabling CLIP LoRA also reduces scores, indicating that frozen RGB-pretrained weights require targeted adaptation for DoLP/AoLP distributions. Finally,

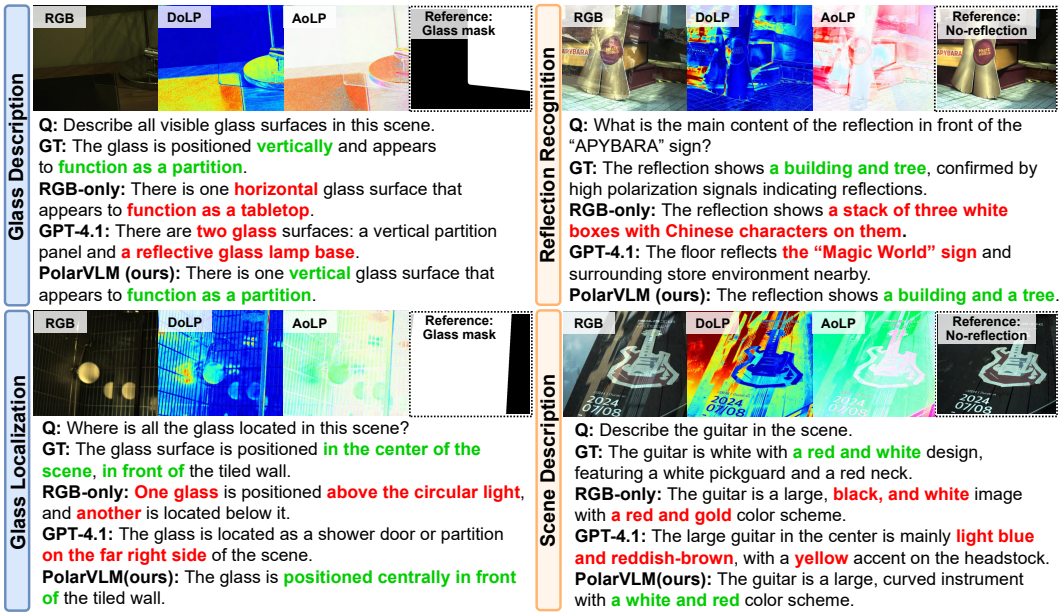

Figure 4: Qualitative comparisons on the PolarVQA test set. We evaluate PolarVLM against the strictly controlled RGB-only LLaVA-1.5-13B (Liu et al., 2024a) baseline and GPT-4.1 (Achiam et al., 2023) under optical ambiguities caused by transparency (left) and specular reflections (right). By grounding its reasoning in polarimetric priors, PolarVLM mitigates the spatial mislocalizations and reflection-induced hallucinations (marked in red), generating physically accurate responses (highlighted in green).

Table 2: Quantitative evaluation results of ablation study.

Configuration	Glass Count	Glass Loc.	Glass Desc.	Scene Desc.	Refl. Recog.	Overall
w/o Stage 1	6.94	5.32	6.45	5.08	3.48	5.46
w/o Norm Layer	7.01	5.41	6.57	4.84	3.76	5.39
w/o CLIP LoRA	6.77	5.25	6.28	5.65	3.36	5.68
w/o RGB Branch	6.49	4.80	6.04	4.47	3.14	4.93
w/o Polar Branch	5.65	4.88	5.77	4.62	3.23	4.85
PolarVLM (Ours)	**7.57**	**5.51**	**6.87**	**5.85**	**4.09**	**6.08**

single-branch variants confirm strict complementarity: the RGB-only branch preserves semantics but fails on physical reasoning, while the polarization-only branch captures structure but loses scene context. Additional ablations and efficiency analysis are in Appendices G and I.

7 Conclusion

We introduced PolarVLM, the first multimodal framework to bridge the semantic-physical gap in open-ended visual question answering, alongside PolarVQA, a comprehensive benchmark dedicated to physics-aware visual reasoning. Through a dual-stream architecture and progressive two-stage training, PolarVLM injects orthogonal physical measurements (*e.g.*, DoLP and AoLP) into the language model while largely retaining RGB semantic priors. Extensive evaluations confirm that PolarVLM resolves severe optical ambiguities, mitigating transparency-induced structural omissions and reflection-induced hallucinations where RGB-only models systematically fail.

Limitations. While PolarVLM demonstrates the potential of polarimetric grounding, our current implementation is bounded by mid-scale open-source backbones, and the dual-stream design introduces moderate computational overhead and a slight trade-off on general-purpose benchmarks. Exploring polarimetric integration with more capable architectures, broader data diversity, and more efficient cross-modal adaptation remains a promising direction.

AI USE STATEMENT

We used GPT-4o-mini as the text synthesizer in the PolarVQA data generation pipeline (§5; Appendix C), where all generated captions and instruction-tuning pairs were programmatically verified against ground-truth physical measurements and further audited by the authors through manual spot-checks before inclusion. For quantitative evaluation, we adopted GPT-4o-mini as the default LLM-as-a-Judge scorer (§6); to mitigate potential evaluator bias, we additionally employed GPT-5.5 and Claude Sonnet 4.5 as independent cross-judges and designed a judge-free multiple-choice protocol (Appendix B). We also used large language models for assistance in proofreading and code debugging; all such outputs were manually reviewed and verified by the authors. We did not use generative AI tools for formulating research ideas, designing experiments, or drawing scientific conclusions. We take responsibility for the final content of this work.

ETHICS STATEMENT

PolarVQA is constructed entirely from publicly available polarimetric datasets (Yao et al., 2025; Mei et al., 2022) and does not involve human subjects, personally identifiable information, or Internet-scraped content. All generated annotations were programmatically verified against ground-truth physical measurements and further audited by the authors through manual spot-checks to ensure factual accuracy. We respect the licenses and terms of use of all source datasets, pretrained models, and codebases employed in this work. A detailed discussion of responsible use and asset compliance is provided in Appendix J.

REPRODUCIBILITY STATEMENT

All architectural details, training hyperparameters, and dataset construction procedures are described in the main text (§3–5) and supplementary materials (Appendices C and D). The PolarVQA benchmark is constructed entirely from publicly available datasets (PolarFree and RGBP-Glass) using the automated pipeline detailed in Appendix C. Evaluation prompts and scoring rubrics are provided in Appendix E. To facilitate reproducibility, we include partial source code and a representative subset of the PolarVQA data in the supplementary material; the complete codebase, model weights, and full dataset will be publicly released upon acceptance.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In *Proc. of European Conference on Computer Vision*, pp. 554–571, 2020.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Xu Cao, Hiroaki Santo, Fumio Okura, and Yasuyuki Matsushita. Multi-view azimuth stereo via tangent space consistency. In *Proc. of Computer Vision and Pattern Recognition*, pp. 825–834, 2023.

Guangcheng Chen, Yicheng He, Li He, and Hong Zhang. PISR: Polarimetric neural implicit surface reconstruction for textureless and specular objects. In *Proc. of European Conference on Computer Vision*, pp. 205–222, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Proc. of Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proc. of Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024c.

Edward Collett. *Field guide to polarization*. SPIE press Bellingham, Washington, 2005.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Proc. of Advances in Neural Information Processing Systems*, pp. 49250–49267, 2023.

Akshat Dave, Yongyi Zhao, and Ashok Veeraraghavan. PANDORA: Polarization-aided neural decomposition of radiance. In *Proc. of European Conference on Computer Vision*, pp. 538–556, 2022.

Valentin Deschaintre, Yiming Lin, and Abhijeet Ghosh. Deep polarization imaging for 3D shape and SVBRDF acquisition. In *Proc. of Computer Vision and Pattern Recognition*, pp. 15567–15576, 2021.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Proc. of Advances in Neural Information Processing Systems*, pp. 10088–10115, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Proc. of Advances in Neural Information Processing Systems*, volume 38, 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proc. of Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Dennis H Goldstein. *Polarized Light*. CRC Press, 2017.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending CLIP to image, text and audio. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 976–980, 2022.

Yufei Han, Heng Guo, Koki Fukai, Hiroaki Santo, Boxin Shi, Fumio Okura, Zhanyu Ma, and Yunpeng Jia. NeRSP: Neural 3D reconstruction for reflective objects with sparse polarized images. In *Proc. of Computer Vision and Pattern Recognition*, pp. 11821–11830, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. of International Conference on Learning Representations*, 2022.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Kailin Jiang, Yuntao Du, Yukai Ding, Yuchen Ren, Ning Jiang, Zhi Gao, Zilong Zheng, Lei Liu, Bin Li, and Qing Li. When large multimodal models confront evolving knowledge: Challenges and explorations. In *Proc. of International Conference on Learning Representations*, 2026a.

Shixin Jiang, Zerui Chen, Jiafeng Liang, Yanyan Zhao, Ming Liu, and Bing Qin. Infrared-LLaVA: Enhancing understanding of infrared images in multi-modal large language models. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 8573–8591, 2024.

Zhiyuan Jiang, Weihao Hong, Xinlei Guan, Tejaswi Dhandu, Miles Q Li, Meng Xu, Kuan Huang, Umamaheswara Rao Tida, Bingyu Shen, Daehan Kwak, et al. LLM-as-judge framework for evaluating tone-induced hallucination in vision-language models. *arXiv preprint arXiv:2604.18803*, 2026b.

Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Depth sensing using geometrically constrained polarization normals. *International Journal of Computer Vision*, 125(1):34–51, 2017.

Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *Proc. of Computer Vision and Pattern Recognition*, pp. 8602–8611, 2020.

Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *Proc. of Computer Vision and Pattern Recognition*, pp. 1750–1758, 2020.

Chenyang Lei, Chenyang Qi, Jiaxin Xie, Na Fan, Vladlen Koltun, and Qifeng Chen. Shape from polarization for complex scenes in the wild. In *Proc. of Computer Vision and Pattern Recognition*, pp. 12632–12641, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. of International Conference on Machine Learning*, pp. 19730–19742, 2023.

Pengteng Li, Yunfan Lu, Pinghao Song, Wuyang Li, Huizai Yao, and Hui Xiong. EventVL: Understand event streams via multimodal large language model. *arXiv preprint arXiv:2501.13707*, 2025.

Junyan Lin, Haoran Chen, Yue Fan, Yingqi Fan, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multi-layer visual feature fusion in multimodal LLMs: Methods, analysis, and best practices. In *Proc. of Computer Vision and Pattern Recognition*, pp. 4156–4166, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proc. of Advances in Neural Information Processing Systems*, pp. 34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proc. of Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *Proc. of European Conference on Computer Vision*, pp. 216–233, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. of International Conference on Learning Representations*, 2019.

Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. Reflection separation using a pair of unpolarized and polarized images. In *Proc. of Advances in Neural Information Processing Systems*, 2019.

Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. Physics-guided reflection separation from a pair of unpolarized and polarized images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2151–2165, 2022.

Youwei Lyu, Lingran Zhao, Si Li, and Boxin Shi. Shape from polarization with distant lighting estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13991–14004, 2023.

Abrar Majeedi, Zhiyuan Ruan, Ziyi Zhao, Hongcheng Wang, Jianglin Lu, and Yin Li. DUALVISION: RGB-infrared multimodal large language models for robust visual reasoning. *arXiv preprint arXiv:2604.18829*, 2026.

Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Glass segmentation using intensity and spectral polarization cues. In *Proc. of Computer Vision and Pattern Recognition*, pp. 12622–12631, 2022.

Taishi Ono, Yuhi Kondo, Legong Sun, Teppei Kurita, and Yusuke Moriuchi. Degree-of-linear-polarization-based color constancy. In *Proc. of Computer Vision and Pattern Recognition*, pp. 19740–19749, 2022.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. In *Proc. of Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802, 2024.

Henry Peters, Yunhao Ba, and Achuta Kadambi. pCON: Polarimetric coordinate networks for neural scene representations. In *Proc. of Computer Vision and Pattern Recognition*, pp. 16579–16589, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. of International Conference on Machine Learning*, pp. 8748–8763, 2021.

Yoav Y Schechner, Srinivasa G Narasimhan, and Shree K Nayar. Instant dehazing of images using polarization. In *Proc. of Computer Vision and Pattern Recognition*, pp. I–I, 2001.

Mingqi Shao, Chongkun Xia, Zhendong Yang, Junnan Huang, and Xueqian Wang. Transparent shape from a single view polarization image. In *Proc. of International Conference on Computer Vision*, pp. 9277–9286, 2023.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. EAGLE: Exploring the design space for multimodal LLMs with mixture of encoders. In *Proc. of International Conference on Learning Representations*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 System Card. *arXiv preprint arXiv:2601.03267*, 2025.

Chaoran Tian, Weihong Pan, Zimo Wang, Mao Mao, Guofeng Zhang, Hujun Bao, Ping Tan, and Zhaopeng Cui. DPS-Net: Deep polarimetric stereo depth estimation. In *Proc. of International Conference on Computer Vision*, pp. 3569–3579, 2023.

Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. CoRRN: Cooperative reflection removal network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(12):2969–2982, 2019.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Xin Wang, Yong Zhang, and Yanchu Chen. Polarized reflection removal with dual-stream attention guidance. *Pattern Recognition*, 157:110945, 2025b.

Junfei Wu, Yue Ding, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. SHARP: Steering hallucination in LVLMs via representation engineering. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 14357–14372, 2025.

Xuesong Wu, Hong Zhang, Xiaoping Hu, Moein Shakeri, Chen Fan, and Juiwen Ting. HDR reconstruction based on the polarization camera. *IEEE Robotics and Automation Letters*, 5(4): 5113–5119, 2020.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proc. of Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 6787–6800, 2021.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *Proc. of Computer Vision and Pattern Recognition*, pp. 26340–26353, 2024.

Mingde Yao, Menglu Wang, King-Man Tam, Lingen Li, Tianfan Xue, and Jinwei Gu. PolarFree: Polarization-based reflection-free imaging. In *Proc. of Computer Vision and Pattern Recognition*, pp. 10890–10899, 2025.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. PointCLIP: Point cloud understanding by CLIP. In *Proc. of Computer Vision and Pattern Recognition*, pp. 8552–8562, 2022.

Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proc. of Computer Vision and Pattern Recognition*, pp. 4786–4794, 2018.

Chu Zhou, Minggui Teng, Yufei Han, Chao Xu, and Boxin Shi. Learning to dehaze with polarization. In *Proc. of Advances in Neural Information Processing Systems*, pp. 11487–11500, 2021.

Chu Zhou, Yufei Han, Minggui Teng, Jin Han, Si Li, Chao Xu, and Boxin Shi. Polarization guided HDR reconstruction via pixel-wise depolarization. *IEEE Transactions on Image Processing*, 32: 1774–1787, 2023.

Chu Zhou, Minggui Teng, Xinyu Zhou, Chao Xu, Imari Sato, and Boxin Shi. Learning to deblur polarized images. *International Journal of Computer Vision*, 133(9):5976–5991, 2025a.

Chu Zhou, Chao Xu, and Boxin Shi. Polarization guided mask-free shadow removal. In *Proc. of the AAAI Conference on Artificial Intelligence*, pp. 10716–10724, 2025b.

Chu Zhou, Yixin Yang, Junda Liao, Heng Guo, Boxin Shi, and Imari Sato. Polarimetric neural field via unified complex-valued wave representation. In *Proc. of International Conference on Computer Vision*, pp. 25660–25669, 2025c.

Shuangfan Zhou, Chu Zhou, Youwei Lyu, Heng Guo, Zhanyu Ma, Boxin Shi, and Imari Sato. PIDSR: Complementary polarized image demosaicing and super-resolution. In *Proc. of Computer Vision and Pattern Recognition*, pp. 16081–16090, 2025d.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Appendix

A General-purpose multimodal benchmarks

To quantify potential catastrophic forgetting from PolarVQA fine-tuning, we evaluate PolarVLM checkpoints on four established multimodal benchmarks: MME (Fu et al., 2025), MMBench (Liu et al., 2024b), MMStar (Chen et al., 2024b), and RealWorldQA. We test across LLaVA-1.5-7B, LLaVA-1.5-13B (Liu et al., 2024a), Qwen3-VL-8B (Bai et al., 2025a), and InternVL3.5-14B (Wang et al., 2025a). In Table 3, "Base" refers to the corresponding pretrained checkpoint prior to any PolarVQA training.

Table 3: General-purpose multimodal evaluation. $\Delta\%$ = (PolarVLM − Base) / Base × 100.

Model	Benchmark	Base	PolarVLM	Δ%
LLaVA-1.5-7B	MME (Perc.)	1510.7	1452.7	−3.8
LLaVA-1.5-7B	MMBench (Circ.)	64.3	62.1	−3.4
LLaVA-1.5-13B	MME (Perc.)	1531.3	1463.5	−4.4
LLaVA-1.5-13B	MMBench (Circ.)	67.7	65.5	−3.2
InternVL3.5-14B	MME (sum)	2398.4	2254.5	−6.0
InternVL3.5-14B	MMStar	70.4	66.9	−5.0
Qwen3-VL-8B	MMStar	70.9	66.4	−6.3
Qwen3-VL-8B	RealWorldQA	71.5	72.9	+2.0

PolarVLM largely retains general-purpose capabilities, with degradation ranging from -3% to -6% across most benchmarks, while RealWorldQA exhibits a modest $+2\%$ improvement. These drops are substantially smaller than the -34% and -17% on MME and MMBench reported by Jiang et al. (2026a) when applying LoRA adaptation on a comparable backbone, suggesting that the dual-stream design effectively limits forgetting by keeping the RGB pathway frozen throughout training.

B Robustness and generalization study

To establish that PolarVLM's improvements are not contingent on a particular evaluator or backbone, we conduct four complementary lines of validation: cross-judge scoring (§B.1), judge-free multiple-choice evaluation (§B.2), human evaluation (§B.3), and multi-backbone generalization (§B.4). Paired statistical analysis is reported in §B.5.

B.1 Cross-judge evaluation

GPT-4o-mini (Hurst et al., 2024) serves as both the text realizer in the PolarVQA pipeline and the default LLM-as-a-Judge scorer, raising a potential self-preference confound (Panickssery et al., 2024). To isolate this, we re-score all model outputs using two independent judges—GPT-5.5 (Singh et al., 2025) and Claude Sonnet 4.5—neither of which participated in PolarVQA construction. The same task-specific rubrics and the complete 1,000-sample test set are used throughout.

Table 4: Cross-judge evaluation on LLaVA-1.5-13B (Liu et al., 2024a). Both RGB-only and Polar models are trained on PolarVQA. Overall is sample-weighted mean. $\Delta\%$ = (Polar − RGB) / RGB × 100.

Judge	Input	Glass Count	Glass Loc.	Glass Desc.	Scene Desc.	Refl. Recog.	Overall	Δ%
GPT-4o-mini	RGB	5.65	4.88	5.77	4.62	3.23	4.85	–
GPT-4o-mini	Polar	7.57	5.51	6.87	5.85	4.09	6.08	+25.4
GPT-5.5	RGB	5.70	4.93	5.82	4.70	3.30	4.92	–
GPT-5.5	Polar	7.65	5.56	6.93	5.90	4.16	6.13	+24.6
Claude Sonnet 4.5	RGB	5.73	5.01	5.78	4.66	3.26	4.91	–
Claude Sonnet 4.5	Polar	7.62	5.58	6.92	5.86	4.15	6.11	+24.4

Tab. 4 shows that all three judges yield consistent relative gains (+24.4–25.4%) on LLaVA-1.5-13B, with per-task rankings preserved across judges. To verify that this consistency extends beyond a

single backbone, we repeat the cross-judge protocol on LLaVA-1.5-7B (Liu et al., 2024a), Qwen3-VL-8B (Bai et al., 2025a), and InternVL3.5-14B (Wang et al., 2025a).

Table 5: Cross-judge Overall scores on three additional backbones. Both RGB-only and Polar models are trained on PolarVQA.

Model	Judge	RGB Overall	Polar Overall	Δ%
LLaVA-1.5-7B	GPT-4o-mini	4.57	5.37	+17.5
LLaVA-1.5-7B	GPT-5.5	4.61	5.42	+17.6
LLaVA-1.5-7B	Claude Sonnet 4.5	4.62	5.41	+17.1
Qwen3-VL-8B	GPT-4o-mini	5.14	6.17	+20.0
Qwen3-VL-8B	GPT-5.5	5.20	6.22	+19.6
Qwen3-VL-8B	Claude Sonnet 4.5	5.21	6.21	+19.2
InternVL3.5-14B	GPT-4o-mini	5.32	6.36	+19.5
InternVL3.5-14B	GPT-5.5	5.38	6.36	+18.2
InternVL3.5-14B	Claude Sonnet 4.5	5.36	6.41	+19.6

As shown in Tab. 5, Polar consistently outperforms the matched RGB model across all 12 backbone–judge combinations, with at most 1.4 percentage points of variation in relative gain for a given backbone. The consistent ranking across three independently developed judges confirms that the observed improvements are not attributable to evaluator-specific biases. We further replace the LLM-based scoring with a judge-free protocol in the following section.

B.2 Judge-free multiple-choice evaluation

To eliminate any LLM judge dependency, we construct **PolarVQA-MCQ**, a 1,000-question, four-option benchmark. Correct options are derived from the ground-truth physical annotations in PolarVQA; distractors are generated by systematic perturbations of key attributes (*e.g.*, altering the glass count, swapping spatial locations, or substituting reflected content). Scoring uses exact letter-match accuracy, requiring no language model at evaluation time. A representative item is shown below.

Q: What is reflected on the black wall near the guitar logo?
A. A blurred cityscape with tall buildings and a blue sky B. People walking in a park
C. The guitar logo itself D. No reflection is visible
Answer: A

We evaluate PolarVQA-MCQ on four backbones: LLaVA-1.5-7B and LLaVA-1.5-13B (Liu et al., 2024a), Qwen3-VL-8B (Bai et al., 2025a), and InternVL3.5-14B (Wang et al., 2025a).

Table 6: PolarVQA-MCQ accuracy (%). Evaluation uses exact letter match without any LLM judge. $\Delta\% = (\text{Polar} - \text{RGB})/\text{RGB} \times 100$.

Model	Input	Glass Count	Glass Loc.	Glass Desc.	Scene Desc.	Refl. Recog.	Overall	Δ%
LLaVA-1.5-7B	RGB	38.9	70.6	40.8	42.6	83.7	48.39	–
LLaVA-1.5-7B	Polar	58.2	78.7	55.3	57.2	90.5	62.34	+28.8
LLaVA-1.5-13B	RGB	46.2	80.6	52.1	50.3	83.7	56.14	–
LLaVA-1.5-13B	Polar	69.3	87.4	63.9	65.2	92.8	70.55	+25.7
Qwen3-VL-8B	RGB	69.7	90.5	89.6	78.9	78.7	80.13	–
Qwen3-VL-8B	Polar	82.4	96.1	92.9	84.4	91.6	87.12	+8.7
InternVL3.5-14B	RGB	66.4	90.3	93.8	81.0	86.3	81.79	–
InternVL3.5-14B	Polar	84.3	95.8	95.4	84.2	96.7	88.02	+7.6

As shown in Tab. 6, Polar outperforms RGB under this judge-free protocol across all four backbones, with gains ranging from +7.6% (InternVL3.5-14B) to +28.8% (LLaVA-1.5-7B). Combined with the cross-judge results (Tables 4 and 5), the Polar-over-RGB advantage holds consistently across evaluator choices and scoring formats.

B.3 HUMAN EVALUATION

We further validate automatic scoring with a human study. Two annotators with physics and computer-vision backgrounds independently score a stratified 200-sample subset covering all five task categories. Annotators follow the same 1–10 rubric used by the LLM judge and are blind to model identity. Inter-annotator disagreements are resolved by averaging. GPT-4o-mini (Hurst et al., 2024) scores are recomputed on this same subset for direct comparison. The study covers all four backbones: LLaVA-1.5-7B, LLaVA-1.5-13B (Liu et al., 2024a), Qwen3-VL-8B (Bai et al., 2025a), and InternVL3.5-14B (Wang et al., 2025a).

Table 7: Human evaluation on the 200-sample subset. $\Delta\%_H$ and $\Delta\%_G$ denote relative gains under human and GPT-4o-mini (Hurst et al., 2024) scoring, respectively.

Model	Input	Glass Count	Glass Loc.	Glass Desc.	Scene Desc.	Refl. Recog.	Human	GPT	$\Delta\%_H$	$\Delta\%_G$
LLaVA-1.5-7B	RGB	5.38	4.76	5.94	4.37	3.00	4.69	4.63	–	–
LLaVA-1.5-7B	Polar	7.00	5.02	6.10	5.40	3.21	5.35	5.40	+14.0	+16.6
LLaVA-1.5-13B	RGB	5.90	4.85	5.78	5.46	3.60	5.12	5.14	–	–
LLaVA-1.5-13B	Polar	7.55	5.23	6.70	5.58	3.72	5.76	5.80	+12.5	+12.8
Qwen3-VL-8B	RGB	6.43	5.25	5.73	5.17	3.40	5.20	5.29	–	–
Qwen3-VL-8B	Polar	7.53	5.54	6.43	5.63	3.88	5.80	5.87	+11.7	+11.0
InternVL3.5-14B	RGB	7.08	4.73	6.51	5.56	3.49	5.47	5.54	–	–
InternVL3.5-14B	Polar	8.15	5.45	6.97	6.25	4.43	6.25	6.21	+14.2	+12.1

Human and GPT-4o-mini Overall scores agree within 0.1 points across all backbones; relative gains range +11.7–14.2% under human scoring versus +11.0–16.6% under GPT-4o-mini. The close agreement confirms that the automatic evaluation protocol is closely aligned with human judgment.

B.4 MULTI-BACKBONE GENERALIZATION

The main evaluation uses a single LLaVA-1.5-13B (Liu et al., 2024a) backbone built on Vicuna-13B and CLIP ViT-L/14 (Radford et al., 2021). To verify that PolarVLM's gains generalize across model families, we apply the same dual-stream formulation and two-stage alignment to three additional architectures spanning different scales and vision encoders: LLaVA-1.5-7B (Liu et al., 2024a), Qwen3-VL-8B (Bai et al., 2025a), and InternVL3.5-14B (Wang et al., 2025a). Table 8 summarizes the configurations. For each variant, the RGB and Polar models share identical training settings; pretrained base weights are frozen, while matched LoRA adapters (Hu et al., 2022) and polarization projectors are optimized.

Table 8: Backbone configurations in the generalization study.

Model	LLM Backbone	Vision Encoder	Family
LLaVA-1.5-13B *(main)*	Vicuna-13B	CLIP ViT-L/14@336	LLaVA
LLaVA-1.5-7B	Vicuna-7B	CLIP ViT-L/14@336	LLaVA
Qwen3-VL-8B	Qwen3-8B	Qwen-ViT	Qwen3-VL
InternVL3.5-14B	Qwen3-14B	InternViT-300M	InternVL

Qwen3-VL and InternVL3.5 use native multimodal architectures with non-CLIP vision encoders. As reported in Tables 4 through 6, adding the polarization stream consistently improves the matched RGB baseline across all four backbones, all three LLM judges, and the judge-free MCQ protocol. The benefit of polarimetric grounding thus generalizes across LLM backbones, vision encoder families, and model scales.

B.5 PAIRED STATISTICAL ANALYSIS

We perform paired analysis on the 1,000 test examples to quantify sampling uncertainty of the reported Overall scores. For each test sample we compute the score difference (PolarVLM minus RGB-only) and apply bootstrap resampling ($B = 1{,}000$) as well as two-sided nonparametric and parametric paired tests.

The two bootstrap confidence intervals for the mean are non-overlapping, and the paired difference CI lies entirely above zero. Both the Wilcoxon signed-rank test and the paired t-test yield $p <$

Table 9: Paired statistical analysis on the 1,000-sample PolarVQA test set (LLaVA-1.5-13B (Liu et al., 2024a), GPT-4o-mini (Hurst et al., 2024) judge, 1,000 bootstrap resamples).

Metric	Value
PolarVLM mean (95% CI)	6.076 [5.870, 6.283]
RGB-only mean (95% CI)	4.850 [4.643, 5.060]
Paired mean difference	+1.226
95% CI of paired difference	[+1.064, +1.386]
Wilcoxon signed-rank p	7.60×10^{-50}
Paired t-test p	7.08×10^{-46}

10^{-45}, indicating that the performance difference between PolarVLM and the RGB-only baseline is statistically significant. The same Polar-over-RGB ranking is preserved under all three independent judges, the judge-free MCQ protocol, and all four backbones (Tables 4–6).

C Details of the PolarVQA generation pipeline

In this section, we detail the automated data generation pipeline used to construct the PolarVQA benchmark from raw pixel-level datasets. As illustrated in Fig. 5, the pipeline comprises three serially connected modules: (a) visual structure extraction, (b) physical quantity analysis, and (c) rigorous text synthesis.

Visual structure extraction. To obtain robust semantic anchors, we employ Florence-2-large (Xiao et al., 2024) for object detection and region description on the RGB images. For reflection scenes (sourced from PolarFree (Yao et al., 2025)), we run Florence-2-large independently on both the original image containing reflections and the reflection-removed reference image. By computing the difference set between these two detection outputs, we strictly isolate the spurious objects that exist exclusively within the specular reflections.

Physical quantity analysis. Subsequently, the physical analysis module bridges the spatial structures with polarimetric cues. For specular reflections, the module localizes reflection regions based on the strict physical prior that reflected light simultaneously exhibits a high DoLP and a substantial RGB domain difference. For transparent object data (sourced from RGBP-Glass (Mei et al., 2022)), the module extracts the spatial attributes and DoLP statistical distributions of each glass instance using the provided COCO-format segmentation masks.

Rigorous text synthesis. Finally, the structured outputs from the previous modules are assembled into standardized text prompts and fed to GPT-4o-mini (Hurst et al., 2024) to generate the final captions and instruction-tuning pairs. To ensure the generation process is fully auditable and free from visual hallucinations, only structured text is submitted to the language model application programming interface (API). We further apply automated rule-based verification to the LLM outputs to filter out formatting errors or unqualified responses. This strictly text-based, physics-constrained synthesis guarantees the high fidelity of the resulting PolarVQA benchmark.

Prompt templates for data generation. To make the text synthesis process transparent and reproducible, we provide the prompt templates used in the four data-generation settings. Specifically, Fig. 6 and Fig. 7 show the caption-generation prompts for Stage 1 (physics-to-semantic alignment), while Fig. 8 and Fig. 9 show the instruction-generation prompts for Stage 2 (joint physics-aware instruction tuning).

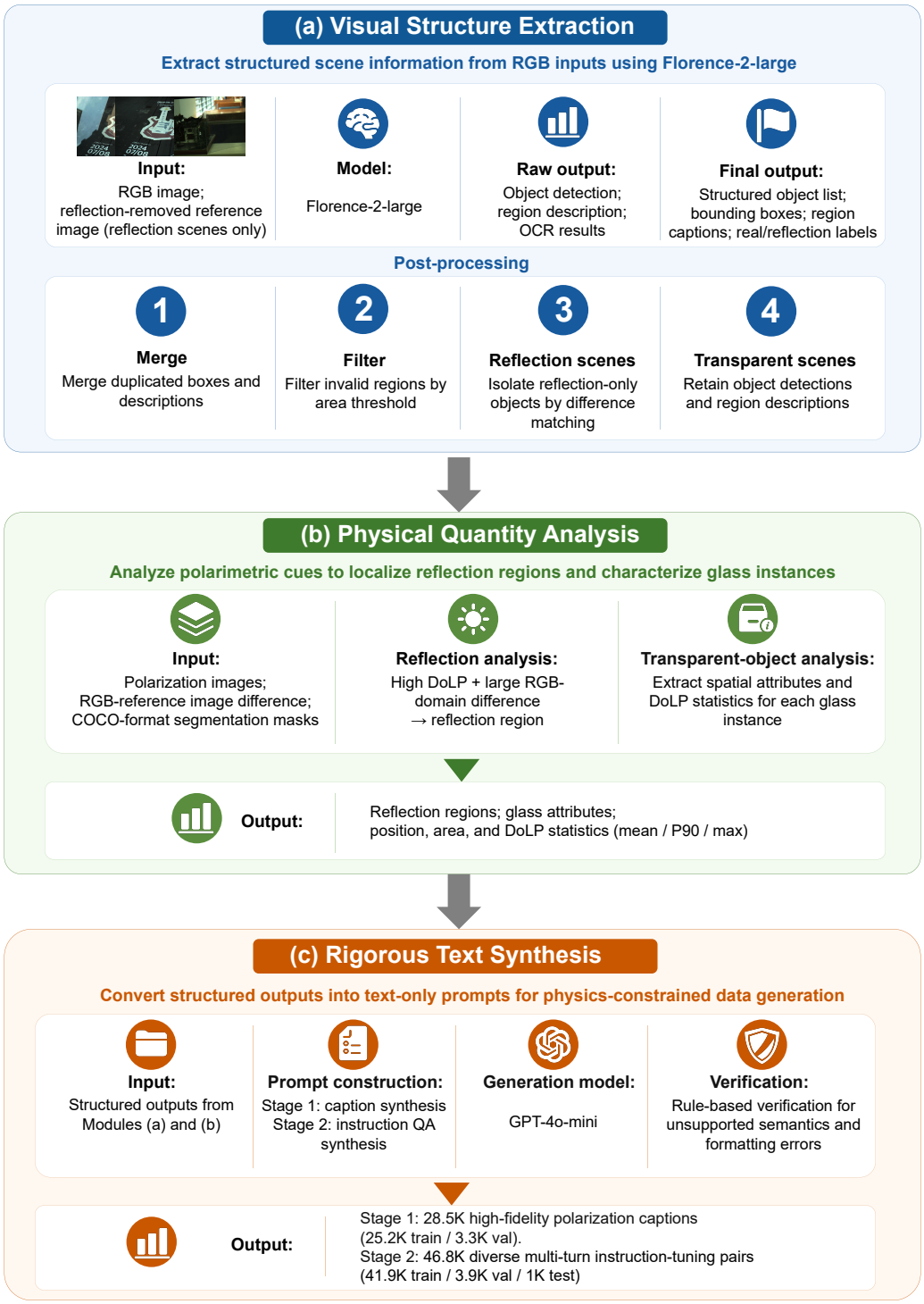

Figure 5: Automated data generation pipeline for PolarVQA. The pipeline extracts visual structures with Florence-2-large (Xiao et al., 2024), analyzes polarimetric cues from reflection and glass data (Yao et al., 2025; Mei et al., 2022), and synthesizes text-only Stage 1 captions and Stage 2 instruction-tuning samples with GPT-4o-mini (Hurst et al., 2024), followed by rule-based verification.

Stage 1 Reflection Caption Generation (Three Variants)

Generation Constraints Prompt:

- Restrict captions to shape, material, surface texture, spatial layout, and reflection cues.
- Prohibit color, text, printed content, social semantics, person details, and unsupported object functions.
- Replace semantic object names with physical descriptions when they require color or text cues.
- Self-check the draft and rewrite violations before output.

Variant 1: Object Shape + Material	Variant 2: Spatial Layout + Geometry	Variant 3: Reflection + Surface Texture
You are a strict caption writer for polarization image-text alignment training. Describe only the physical shapes and materials of objects visible in a polarization image. Do not rely on color, text, or high-level semantics. Focus on: (1) describing the main objects by shape, such as rectangular, cylindrical, curved, flat, circular, triangular, or irregular; (2) identifying material when supported by polarization cues, such as metal, wood, glass, plastic, fabric, stone, concrete, or paper; (3) referring to people only as "human silhouette(s)"; (4) using a generic physical description when an object cannot be reliably named from shape and material alone; (5) mentioning reflection only briefly by its location and intensity, or stating "No significant surface reflection" if absent. Output a concise English caption in 10--30 words only.	You are a strict caption writer for polarization image-text alignment training. Describe only the spatial layout and geometric relationships visible in a polarization image. Do not use color, text, or high-level semantics. Focus on: (1) describing object locations, such as foreground/background, left/right/center, and upper/lower regions; (2) describing geometric relationships, such as aligned, stacked, overlapping, side-by-side, perpendicular, or receding; (3) using physical terms such as outline, contour, silhouette, edge, plane, surface, and structure; (4) referring to people only as "human silhouette(s)"; (5) mentioning reflection only briefly by its location and intensity, or stating "No significant surface reflection" if absent. Output one concise English caption in 10--30 words.	You are a strict caption writer for polarization image-text alignment training. Describe only the surface texture and reflection characteristics visible in a polarization image. Do not use color, text, or high-level semantics. Focus on: (1) describing surface textures, such as smooth, rough, glossy, matte, corrugated, slatted, granular, polished, or worn; (2) describing reflection location and intensity when reflection is present; (3) stating "No significant surface reflection" when no reflection is observed; (4) briefly mentioning the dominant surface type for context; (5) referring to people only as "human silhouette(s)"; (6) avoiding detailed object enumeration and focusing on scene-level surface properties. Output one concise English caption in 10--30 words.

Verification Prompt:

You are a strict quality checker for polarization image captions.
Check whether the caption contains only information observable in polarization imagery. Focus on:

- allowing only physically grounded content: shape, material, spatial layout, surface texture, and reflection location/intensity;
- removing forbidden content: colors, readable text, printed content, social semantics, person details, and action descriptions;
- replacing scene-specific function names with generic physical descriptions, such as "metallic box" or "flat glass screen";
- rewriting food items or merchandise as generic descriptions such as "small objects" or "irregular objects";
- returning the caption unchanged if it is already clean, otherwise outputting a corrected version.

Example:

User:

GT Caption:
A customer buys pastries at a bright yellow bakery counter while another person waits near a red fire extinguisher by the glass display case.
Detected Objects:
bakery counter [0.12, 0.55, 0.88, 0.92]
fire extinguisher [0.02, 0.30, 0.08, 0.65]
glass display case [0.15, 0.40, 0.85, 0.70]
person [0.35, 0.25, 0.55, 0.85]
Reflection Info:
Surface reflection detected in the lower region, intensity: moderate.

Assistant:

Caption:
A human silhouette stands before a long rectangular counter and a large rectangular glass structure; moderate surface reflection appears in the lower area.
(Check: The output removes color, food, and object-function details while preserving the given reflection cue.)

Figure 6: Prompt template for reflection caption generation in Stage 1. The template contains shared generation constraints, three caption variants, a verification prompt, and an illustrative example for producing polarization-grounded reflection captions.

Stage 1 Glass Caption Generation (Three Variants)

Generation Constraints Prompt:

- Each caption must mention both glass surfaces and at least one non-glass object described by shape and material.
- Each glass surface must include its polarization intensity: strong, moderate, weak, or very weak.
- When glass polarization is higher than the background, the caption must explicitly describe the relative contrast.
- Prohibit colors, readable text, numbers, coordinates, square brackets, social semantics, and unsupported object functions.
- If many glass surfaces are present, describe the 2–3 largest ones and summarize the rest as "several smaller glass surfaces".

Variant 1: Object-First Scene Overview	Variant 2: Spatial Arrangement	Variant 3: Polarization Contrast
You are a concise caption writer for polarization image-text alignment training. Describe the scene by starting from the most prominent non-glass object, then naturally introduce the glass surfaces and their polarization. Focus on: (1) starting with a non-glass object described by shape and material; (2) introducing glass after the object using natural connectors such as "with", "near", or "alongside"; (3) describing glass polarization as a physical attribute, such as weak or moderate polarization; (4) avoiding captions that start directly with "Glass" or "A glass"; (5) writing one flowing scene-level sentence.	You are a concise caption writer for polarization image-text alignment training. Describe how glass surfaces and surrounding objects are spatially arranged in the scene. Focus on: (1) using spatial relationships such as in front of, behind, between, beneath, above, adjacent to, or to the left/right of; (2) linking glass and non-glass objects through position rather than listing them separately; (3) describing glass location, size, and polarization intensity; (4) using at least two spatial expressions when possible; (5) writing one concise caption centered on spatial layout.	You are a concise caption writer for polarization image-text alignment training. Describe glass surfaces by emphasizing their polarization contrast with surrounding non-glass surfaces. Focus on: (1) starting with glass surfaces and their polarization intensity; (2) comparing glass with surrounding surfaces using expressions such as "contrasting with", "standing out from", or "against"; (3) mentioning whether non-glass surroundings show minimal, weak, or no polarization; (4) emphasizing that glass differs from diffuse or unpolarized background regions; (5) avoiding a purely spatial description or object list.

Verification Prompt:

You are a strict quality checker for polarization image captions.
Check whether the caption contains only information observable from polarization imagery.
Focus on:

- removing forbidden content, including colors, readable text, printed content, social actions, subjective descriptions, and unsupported object functions;
- replacing scene-specific object names with generic physical descriptions, such as "metallic box", "flat glass screen", or "rectangular container";
- ensuring every glass surface has a polarization intensity word: strong, moderate, weak, or very weak;
- removing contradictory descriptions, such as strong and weak polarization assigned to the same surface;
- keeping background polarization descriptions consistent throughout the caption;
- deleting vague filler phrases and retaining only shape, material, layout, surface texture, and polarization cues.

If the caption is clean, return it unchanged. Otherwise, output the corrected version only.

Example:

User:

Captions:
Shoppers browse pastries behind a blue glass bakery counter; a red fire extinguisher stands beside the entrance display.
Boxes — Objects:
bakery counter bbox=[0.11, 0.54, 0.89, 0.93] → position: lower-center
fire extinguisher bbox=[0.02, 0.28, 0.09, 0.72] → position: middle-left
human silhouette bbox=[0.42, 0.31, 0.58, 0.88] → position: middle-center
Boxes — Glass:
Glass 1: location=center foreground, size=large, shape=flat rectangular, area=18.0% of image, moderate polarization (p90=0.242, mean=0.118, max=0.389), noticeably higher than background, bbox=[0.14, 0.52, 0.86, 0.91] → lower-center
Plus 2 additional smaller glass surface(s) with weak polarization (omitted for brevity)
Spatial relationships: flat glass panel in front of shelving structure
Total glass coverage: 24.5% of image
Polarization summary:
Overall scene polarization: moderate (mean=0.082, p90=0.156, max=0.412)
Background (non-glass) polarization: low (mean=0.031, p90=0.068)
Pixels with notable polarization (DoLP>0.15): 12.4% of image
KEY CONTRAST: Glass polarization is NOTICEABLY HIGHER than the non-glass background. Glass surfaces can be differentiated from surroundings by their polarization.

Assistant:

Caption:
A long rectangular metallic counter runs across the lower scene with a human silhouette nearby; a large flat glass surface in the center shows moderate polarization that stands out from the weakly polarized surroundings.
(Check: No colors, text, or numerical coordinates appear in the output; glass polarization is stated in qualitative terms and aligned with the "noticeably higher than background" cue from DoLP, consistent with the polarization-caption constraints.)

Figure 7: Prompt template for glass caption generation in Stage 1. The template guides the generation of polarization-grounded captions for transparent-object scenes while suppressing unsupported RGB-only semantics such as colors, readable text, and ambiguous object functions.

Stage 2 Reflection QA Pair Generation (Three Task Types)

Generation Constraints Prompt:

- Compare the CLEAN caption and the OBSERVED caption to distinguish real objects from reflection artifacts.
- Content appearing in the CLEAN caption corresponds to real physical objects.
- Content appearing only in the OBSERVED caption is treated as reflection-induced content.
- Use detected objects/text and bounding boxes only as supporting evidence for object identity and spatial anchors.
- Do not output raw coordinates or mention data-source terms such as “GT”, “RGB”, “clean caption”, or “observed caption”.
- All generated answers must be evidence-based and should not invent attributes absent from the captions or detections.

Task A: Scene Description	Task B: Reflection Recognition	Task C: Counterfactual Reasoning
Goal: Generate QA pairs about true physical objects revealed after reflection removal. Focus on: (1) prioritizing objects that appear in the clean caption but are missing, obscured, or distorted in the observed caption; (2) asking about important real objects, their attributes, or spatial relationships; (3) mixing question types, including description, attribute, and relative-position questions; (4) naming a concrete object in each question; (5) answering using only facts from captions or detections. Output: Generate multiple independent QA pairs, up to five in total. Each answer should be one concise sentence and should not mention polarization, reflection removal, or image-processing terminology.	Goal: Generate one QA pair that asks the model to identify what is reflected on a surface. Focus on: (1) comparing the two captions word by word to find content present only in the observed caption; (2) treating only observed-only content as reflection content; (3) never describing objects appearing in both captions as reflections; (4) preferring concrete reflected objects such as buildings, windows, trees, or silhouettes over vague glare; (5) designing the question with a detailed surface or nearby-object anchor without revealing the reflected object. Answer requirements: Describe the reflected content, state where it appears using a natural scene anchor, and explain that it is confirmed by high polarization or DoLP signals.	Goal: Generate a deceptive existence question that teaches the model not to treat reflection artifacts as real objects. Focus on: (1) selecting one reflected target that appears in the observed caption but not in the clean caption; (2) prioritizing concrete reflected objects over vague background differences; (3) asking whether the target physically exists in the scene; (4) avoiding questions such as “Can you see ...?”, because reflections may be visible; (5) using nearby real objects or surfaces as anchors only when supported by the captions. Answer requirements: The answer must start with “No.” It should explain that the apparent object is a surface reflection rather than a real physical object, supported by high polarization or DoLP evidence.

Quality Control Prompt:

Check each generated QA pair before output.
Focus on:

- ensuring that every answer is supported by the provided captions, detections, or reflection analysis;
- ensuring that objects appearing in both clean and observed captions are treated as real objects, not reflections;
- ensuring that reflection targets are selected only from observed-only content;
- removing raw coordinates and data-source terms such as “clean”, “observed”, “GT”, or “RGB”;
- using natural spatial anchors instead of coordinate-derived phrases;
- returning only valid JSON with no markdown or extra explanation.

Example:

User:

Captions — Clean Scene (reflections removed):
A convenience store aisle with a red metallic bench against the wall and dark grey tiled flooring; fluorescent tube lighting runs along the ceiling.

Captions — Observed Scene (with reflections/glare):
Same aisle with the tiled floor showing a mirrored view of a multi-story building facade and illuminated storefront lettering beside the bench.

Boxes — Clean:
bench: [0.12, 0.42, 0.38, 0.72]
text "EXIT": [0.02, 0.08, 0.09, 0.14]

Boxes — Observed:
bench: [0.12, 0.42, 0.38, 0.72]
building: [0.18, 0.62, 0.92, 0.91]
text "EXIT": [0.02, 0.08, 0.09, 0.14]

Polarization Reflection Analysis:
Total reflection coverage: 14.2% of the image.
Region 1: located at the lower-center of the image, moderate intensity, covering 12.8% of image area, clear structural contours, structured reflection with identifiable shapes or edges, bounding box [0.17, 0.61, 0.93, 0.93].

Reflection-Only Objects (in observed but NOT in clean):
building: [0.18, 0.62, 0.92, 0.91]

Specific Requirement:
Follow the six steps: list objects from clean vs observed; treat items present only in observed as reflections; prefer reflection-only detections; pick the main reflected content; phrase the question with a detailed anchor (surface + nearby real cues), without naming the reflected object in the question; answer with what is reflected, grounded in nearby scene wording (no raw bbox directions), and explain confirmation via high DoLP / polarization.

Assistant:
Question: What is reflected on the dark grey tiled floor near the red metallic bench?
Answer: The reflection shows a multi-story building facade on the floor tiles beside the bench. High polarization (DoLP) in that region matches specular reflection from the glossy tile, so it is glare-mediated imagery—not a physical structure lying on the floor.

Figure 8: Prompt template for reflection instruction generation in Stage 2. The template is used to synthesize physics-aware question-answer pairs for reflection-related tasks, including scene description, reflection recognition, and counterfactual reasoning.

Stage 2 Glass QA Pair Generation (Four Task Types)

Generation Constraints Prompt:

- Analyze the RGB image together with glass metadata, including the number of glass surfaces, coarse locations, shapes, spatial relationships.
- Use the metadata as the factual basis for glass count and position, but describe answers using only objects and surfaces visible in the image.
- Do not hallucinate scene objects or unsupported anchors, such as unseen desks, signs, corridors, or fixtures.
- Do not mention data-source terms such as “metadata”, “annotation”, “COCO”, “bbox”, “ground truth”, area ratio, or percentage.
- Do not add polarization-related boilerplate such as DoLP, AoLP, or polarimetric cues.
- Keep each answer concise, usually within one or two sentences.

Task A: Glass Detection	Task B: Glass Counting	Task C: Glass Localization	Task D: Glass Description
Goal: Generate a presence-style QA pair asking whether glass exists near a specific visible object. Focus on: (1) choosing a small, specific, and identifiable anchor near a glass surface; (2) prioritizing unique anchors such as brand logos, text signs, distinctive objects, furniture, fixtures, or architectural details; (3) avoiding broad anchors such as “wall”, “floor”, “building”, “surface”, or abstract positions; (4) never using the glass itself as the anchor; (5) making the question sound like a natural visual inquiry. Answer requirements: Answer only “YES”. Do not add extra descriptions.	Goal: Generate a QA pair asking how many glass surfaces are visible in the scene. Focus on: (1) using the metadata as the ground-truth source for the number of glass surfaces; (2) avoiding contradiction with the annotated glass count; (3) not estimating area ratios or percentages; (4) not adding unsupported glass surfaces; (5) keeping the answer direct and concise. Answer requirements: State the correct number of glass surfaces clearly.	Goal: Generate a QA pair about where the glass surfaces are located. Focus on: (1) using the exact provided localization question when required; (2) describing each glass surface with visible scene-object anchors; (3) independently anchoring every glass surface to its own real object or scene structure; (4) never using another glass surface as a positional reference; (5) avoiding phrases such as “the second glass is next to the first glass” or “another glass is beside it”. Answer requirements: Describe the location of each glass surface using specific visible anchors, such as furniture, signs, fixtures, or architectural structures. Keep the answer to one or two sentences.	Goal: Generate a QA pair describing the observable properties of all visible glass surfaces. Focus on: (1) stating the number of glass surfaces when appropriate; (2) always mentioning orientation, such as vertical, horizontal, or angled; (3) describing relative size differences when multiple glass surfaces are present; (4) optionally inferring function when visually supported, using cautious phrasing such as “appears to be” or “likely serves as”; (5) avoiding filler phrases such as “reflecting light”, “smooth texture”, “clear and transparent”, or unsupported function claims. Answer requirements: Use one or two concise sentences. Do not directly copy metadata shape terms; describe only what is visually observable.

Quality Control Prompt:

Check each generated QA pair before output.
Focus on:

- ensuring the glass count and location do not contradict the metadata;
- ensuring every referenced anchor is actually visible in the RGB image;
- ensuring glass detection uses a specific non-glass anchor;
- ensuring glass localization does not use glass-to-glass references;
- removing raw coordinates, exact percentages, area ratios, and data-source terms;
- removing unsupported objects, unsupported functions, and generic filler phrases;
- returning only a valid JSON array in the required format.

Example:

User:

Captions:

An indoor mall walkway with grey square floor tiles; ahead is a boutique façade framed by pale stone pillars; a bright rectangular TISSOT sign hangs centered above paired entrance panels; to the left of the doorway sits a red cylindrical fire extinguisher on a low metallic stand; twin tall transparent panels form the doors and a narrower horizontal band runs above them.

Glass layout (metadata):

Glass count: 3
Glass 1: location = left door panel
Glass 2: location = right door panel
Glass 3: location = transom band above the doors
Spatial relationships:
The left and right door panels are vertically aligned on either side of the central entrance gap.
The transom spans above both door panels between the pillars.

Specific Requirement:

Treat the caption as the only scene inventory: mention only objects or structures supported there; use the metadata as ground truth for how many glass regions exist and their coarse locations, without citing “metadata” or percentages. Produce three QA types—(1) glass_detect: ask whether glass appears near one small, specific anchor visible in the caption (answer YES); (2) glass_location: answer the fixed question “Where is all the glass located in this scene?” so each glass is anchored to a different non-glass scene object (never relative to another glass); (3) glass_describe: answer the fixed prompt “Describe all visible glass surfaces in this scene.” with orientation and relative extent among panels, without banned size words like “large” or “small”. Keep answers to one or two short sentences.

Assistant:

Question: Is there glass near the illuminated TISSOT sign above the entrance?
Answer: YES.
Question: Where is all the glass located in this scene?
Answer: One vertical panel forms the left storefront door beside the red cylindrical extinguisher, another vertical panel forms the right door, and a horizontal band runs above both panels beneath the TISSOT sign between the stone pillars—each tied to a different landmark, not to each other.
Question: Describe all visible glass surfaces in this scene.
Answer: There are three: two tall vertical panels that serve as the paired entrance doors, and one horizontal panel angled flat above them; the side panels match in height while the upper band is shorter vertically and stretches across the doorway span.

Figure 9: Prompt template for glass instruction generation in Stage 2. The template is used to synthesize instruction-tuning pairs for glass-related tasks, including glass detection, glass counting, glass localization, and glass description.

D COMPREHENSIVE HYPERPARAMETER CONFIGURATION

Architecture settings. The CLIP ViT-L/14 encoder (Radford et al., 2021) operates at a 336×336 resolution and produces 576 visual tokens with a feature dimension of 1024. The language-model hidden size is 5120, and the maximum sequence length is 2048. The polarization projector is a two-layer MLP, $1024 \to 5120 \to 5120$, followed by LayerNorm.

Stage 1: physics-to-semantic alignment. Stage 1 optimizes only the polarization branch in a polar-only setting, where the RGB branch is bypassed. The trainable modules include the polarization projector and the low-level input layers of the polarization CLIP encoder, including patch embedding, class token embedding, positional embedding, and the pre-Transformer LayerNorm; the LLM, RGB branch, and remaining CLIP Transformer layers are frozen. This stage contains approximately 32.7M trainable parameters. We train for 5 epochs using AdamW (Loshchilov & Hutter, 2019), with learning rates of 5×10^{-4} for the projector and 5×10^{-5} for the unfrozen CLIP layers. The warmup ratio is 3%, the weight decay is 0.05, and the effective batch size is 32. On two NVIDIA RTX 5880 Ada GPUs (48GB VRAM each), this stage takes approximately 10 hours.

Stage 2: joint physics-aware instruction tuning. Stage 2 applies QLoRA adapters (Dettmers et al., 2023) with $r = 64$, $\alpha = 128$, and dropout 0.1 to the LLM `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` matrices, and to the `q_proj`, `k_proj`, and `v_proj` matrices of all 24 polarization CLIP Transformer layers. The second linear layer of the polarization projector is frozen, while the first remains trainable; the polarization CLIP embedding layers, RGB CLIP encoder, and RGB projector are frozen. This stage contains approximately 265M trainable parameters. We train for 3 epochs, about 15.7K optimization steps, using learning rates of 2×10^{-4}, 1×10^{-4}, and 2×10^{-5} for LLM LoRA, the polarization projector, and polarization CLIP LoRA, respectively. The warmup ratio is 3%, the weight decay is 0.05, and the effective batch size is 8. On two NVIDIA RTX 5880 Ada GPUs (48GB VRAM each), this stage takes approximately 20 hours.

E EVALUATION PROTOCOL DETAILS

Test-set curation and task selection. While our instruction tuning phase utilizes seven distinct task categories, we intentionally curate the test set to focus exclusively on five complex, open-ended reasoning tasks. Binary classification tasks (*e.g.*, *glass detection*) are excluded from the evaluation split to prevent the overall metrics from being skewed by trivial yes/no guessing biases. Similarly, the *counterfactual reasoning* subset serves explicitly as a training-time regularization mechanism to penalize reflection-induced hallucinations, rather than a standalone test metric.

Rationale for LLM-as-a-Judge. Because PolarVQA expects open-ended, highly descriptive natural language responses, traditional metrics like exact string matching are inadequate for comprehensively assessing semantic and physical correctness. Consequently, following established rigorous protocols (Wu et al., 2025; Jiang et al., 2026b), we employ the LLM-as-a-Judge paradigm, which has proven highly reliable for evaluating factual consistency and hallucination severity.

Details on computing the overall score. To accurately reflect the model's performance proportional to the test data distribution, the overall score is calculated as a sample-weighted average. Let N_t denote the number of test samples for task t, and $S_{t,i}$ denote the LLM-assigned score (on a 1–10 scale) for the i-th sample in task t. The overall metric is formally computed as:

$$\text{Overall Score} = \frac{\sum_{t \in \mathcal{T}} \sum_{i=1}^{N_t} S_{t,i}}{\sum_{t \in \mathcal{T}} N_t}, \tag{6}$$

where $\mathcal{T}$ represents the set of the five evaluated cognitive tasks.

Evaluation prompts and stability. As shown in Fig. 10, we use task-specific judging prompts for five tasks: *scene description*, *reflection recognition*, *glass counting*, *glass description*, and *glass localization*. All prompts use the same text-only template, including the question, reference answer, model prediction, and scoring criteria, while emphasizing task-specific aspects such as numerical correctness, spatial consistency, glass attributes, scene semantics, and polarimetrically verified reflected content. To ensure stable and comparable evaluation, all models are evaluated on the same fixed PolarVQA test split under temperature-zero or deterministic decoding where supported, and the LLM-as-a-Judge scoring is repeated and averaged using the same prompts across all methods.

Evaluation Prompt (Scene Description)

You are evaluating an AI assistant's answer to a visual question about objects or content in an image.
Please compare the Model Answer with the Ground Truth Answer.
Rate the response based on the following criteria:
(1) Accuracy: Does the model answer align factually with the Ground Truth?
(2) Relevance: Does the model address the specific question asked?
(3) Detail: Does the model provide an appropriate level of detail without being too vague or hallucinating extra content?
Score on a scale of 1 to 10, where higher is better. A model answer that contradicts the Ground Truth should receive a low score from 1 to 4. A model answer that is correct but less detailed than the Ground Truth can receive a score from 5 to 7. A model answer that matches the Ground Truth well should receive a score from 8 to 10.
Please first output a single line containing only one numerical value indicating the score from 1 to 10.

Evaluation Prompt (Glass Counting)

You are evaluating an AI assistant's answer about counting the number of glass pieces or panels in a scene.
Evaluation criteria, in order of importance:
(1) Count accuracy: Does the model give the correct number of glass pieces? An exact match is ideal. Being off by 1 is acceptable and should usually receive a score around 5 to 6. Being off by 2 or more should receive a very low score from 1 to 3.
(2) Understanding: Does the model demonstrate that it is counting glass specifically?
(3) Confidence: Does the model provide a clear and definitive count rather than a vague or hedging answer?
Score on a scale of 1 to 10:
1 to 3: Wrong count, off by 2 or more, or fails to identify glass at all.
4 to 5: Partially correct, off by 1, or identifies glass but gives an unclear count.
6 to 7: Correct count but minimal reasoning.
8 to 10: Correct count with clear understanding.
Please first output a single line containing only one numerical value indicating the score from 1 to 10.

Evaluation Prompt (Glass Description)

You are evaluating an AI assistant's answer describing the visual characteristics of glass in a scene, such as type, appearance, reflections, transparency, shape, or context.
The Ground Truth answer provides a reference description based on expert analysis.
Evaluation criteria, in order of importance:
(1) Descriptive accuracy: Does the model correctly describe the glass characteristics mentioned in the Ground Truth?
(2) Key details: Does the model mention important features such as transparency, reflection, size, shape, or surrounding context?
(3) Factual alignment: Does the model avoid contradicting the Ground Truth?
(4) Detail level: Does the model provide appropriate detail without hallucinating non-existent features?
Score on a scale of 1 to 10, where higher is better. A model answer that contradicts key aspects of the Ground Truth should receive a low score from 1 to 4.
Please first output a single line containing only one numerical value indicating the score from 1 to 10.

Evaluation Prompt (Reflection Recognition)

You are evaluating an AI assistant's answer about reflection phenomena on surfaces, such as what is reflected on a glass window, wall, or floor. The question asks the model to identify or describe reflected content.
The Ground Truth answer is derived from specialized imaging analysis, including polarization imaging, which can reliably detect and characterize surface reflections. The Ground Truth should be treated as the definitive correct answer.
Evaluation criteria, in order of importance:
(1) Content accuracy: Does the model correctly identify the reflected content described in the Ground Truth? The model does not need to mention polarization or DoLP, but should correctly describe what is reflected.
(2) Factual alignment: Does the model agree with the Ground Truth on whether a reflection exists and what it contains?
(3) Detail and completeness: Does the model provide relevant details consistent with the Ground Truth?
Please first output a single line containing only one numerical value indicating the score from 1 to 10.

Evaluation Prompt (Glass Localization)

You are evaluating an AI assistant's answer about the location of glass in a scene.
The Ground Truth answer is derived from specialized imaging analysis, including polarization imaging and segmentation masks, which can reliably detect and locate transparent glass surfaces.
Important evaluation guidelines:
The Ground Truth and Model Answer may use different anchor objects to describe the same location. For example, the Ground Truth may say “near the potted plant”, while the Model Answer may say “beside the wooden shelf”. These may refer to the same spatial region. Do not penalize the model for using different anchor objects if the spatial region is plausibly the same.
Focus on broad spatial region matching, such as left, right, center, top, or bottom of the scene, rather than exact anchor object matching. Since glass is transparent and often nearly invisible, any reasonable attempt to locate it should receive partial credit.
Evaluation criteria, in order of importance:
(1) Glass count consistency: Does the model identify a similar number of glass pieces as the Ground Truth? If the Ground Truth mentions two pieces and the model says one or three, deduct points.
(2) Approximate spatial region: Does the model place the glass in roughly the same area of the scene as the Ground Truth?
(3) Specificity: Does the model provide some spatial context, even if the anchor objects differ from the Ground Truth?
Score on a scale of 1 to 10:
1 to 2: Completely wrong region or fails to mention glass at all.
3 to 4: Wrong count and wrong region, but at least attempts to describe the location.
5 to 6: Correct count or correct approximate region.
7 to 8: Correct count and approximately correct region.
9 to 10: Correct count, correct region, and clear spatial description.
Please first output a single line containing only one numerical value indicating the score from 1 to 10.

Template	Example1: Glass Counting	Example2: Scene Description
User: Ground Truth Answer: {gt_answer} Model Answer: {model_answer} **Assistant:** Score (1-10)	**User:** Ground Truth Answer: There are 10 pieces of glass. Model Answer: There are 8 pieces of glass. **Assistant:** 3.0	**User:** Ground Truth Answer: The background is black. Model Answer: The background is a light blue. **Assistant:** 1.0

Figure 10: Task-specific evaluation prompts used by GPT-4o-mini (Hurst et al., 2024). The evaluator compares the model prediction with the reference answer using text-only inputs and assigns a score on a 1–10 scale.

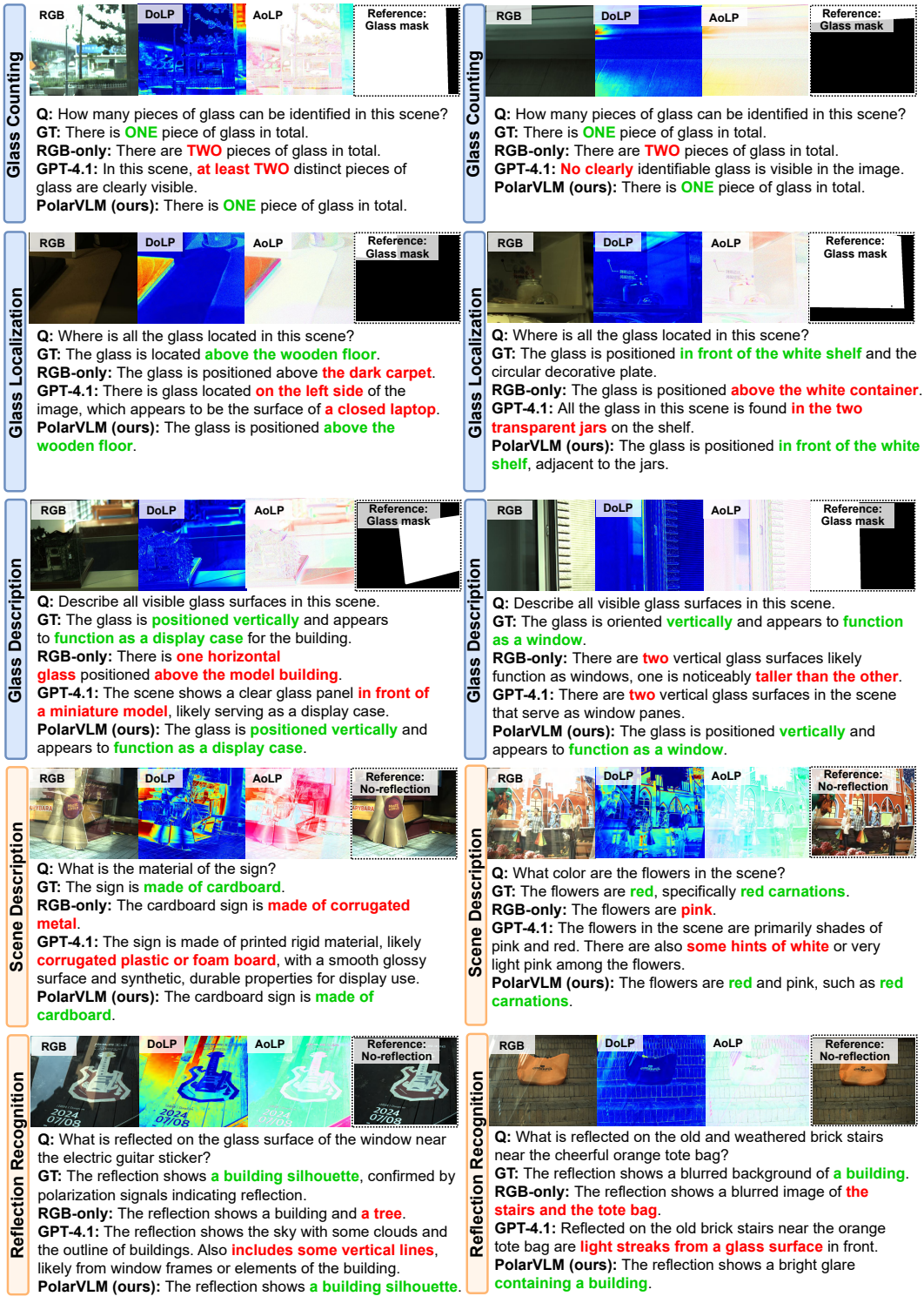

Figure 11: Additional qualitative comparisons on the PolarVQA test set. The examples cover glass counting, glass localization, glass description, scene description, and reflection recognition. Green text denotes responses that are correct or closely aligned with the ground truth, while red text marks incorrect or hallucinated content.

F ADDITIONAL QUALITATIVE RESULTS

To further illustrate the behavior of PolarVLM under diverse optical ambiguities, Fig. 11 provides additional qualitative comparisons on the PolarVQA test set. The examples cover five evaluation tasks, including *glass counting*, *glass localization*, *glass description*, *scene description*, and *reflection recognition*. Compared with the RGB-only LLaVA-1.5-13B (Liu et al., 2024a) baseline and GPT-4.1 (Achiam et al., 2023), PolarVLM more consistently aligns its responses with the ground-truth annotations by leveraging polarimetric cues, especially when transparent surfaces or reflected content are difficult to infer from RGB appearance alone.

G ADDITIONAL ABLATION STUDIES

Table 10: Ablation of RGB–polarization fusion strategies. We compare direct sequence concatenation with representative fusion paradigms from prior multimodal VLMs.

Fusion strategy	Glass Count	Glass Loc.	Glass Desc.	Scene Desc.	Refl. Recog.	Overall
Local Attention	6.64	5.02	5.99	4.49	3.40	5.00
Channel Concatenation	6.64	5.36	6.36	4.96	3.33	5.33
Element-wise Addition	6.42	5.07	6.15	4.72	3.81	5.14
Sequence Concat. (Ours)	**7.57**	**5.51**	**6.87**	**5.85**	**4.09**	**6.08**

Fusion strategy ablation. Tab. 10 compares our straightforward sequence concatenation against representative multimodal fusion paradigms. We first evaluate Multi-scale Localized Cross-Attention (Majeedi et al., 2026), which forces early fusion of spatially aligned neighboring tokens before LLM decoding. While effective for substituting degraded modalities (*e.g.*, RGB–Infrared), it severely underperforms here. This indicates that polarimetric cues provide orthogonal, physical evidence (*e.g.*, surface material properties) rather than mere local visual replacements, making early spatial fusion suboptimal. Similarly, intermediate fusion techniques, such as Channel Concatenation (Shi et al., 2025) and Element-wise Addition (Lin et al., 2025), compress RGB and polarization features into shared token representations. This premature merging inevitably scrambles the distinct statistical signatures of each modality. In contrast, our direct sequence concatenation preserves independent, uncompressed tokens for both branches. By delegating the fusion mechanism entirely to the LLM's deep self-attention layers, the model dynamically cross-references appearance semantics and physical priors based on the specific reasoning context, yielding the highest performance across all cognitive tasks.

Table 11: Ablation of polarization encoding strategies. All variants use the same architecture and fusion strategy.

Encoding	Glass Count	Glass Loc.	Glass Desc.	Scene Desc.	Refl. Recog.	Overall
S_0+Stokes	6.77	5.30	6.26	4.85	3.61	5.30
DoLP-Coupled Angular	6.78	5.09	6.31	4.96	3.73	5.35
S_0+DoLP+AoLP	6.64	5.32	6.05	5.69	**4.44**	5.74
Decoupled (Ours)	**7.57**	**5.51**	**6.87**	**5.85**	4.09	**6.08**

Encoding ablation. We compare four polarization encoding strategies, defined as:

$$S_0\text{+Stokes:} \quad \mathbf{X}_{\text{pol}} = [S_0,\ S_1,\ S_2], \tag{7}$$

$$\text{DoLP-Coupled Angular:} \quad \mathbf{X}_{\text{pol}} = [P,\ P \cdot \cos(2\Phi),\ P \cdot \sin(2\Phi)], \tag{8}$$

$$S_0\text{+DoLP+AoLP:} \quad \mathbf{X}_{\text{pol}} = [S_0,\ P,\ \Phi], \tag{9}$$

$$\text{Decoupled (Ours):} \quad \mathbf{X}_{\text{pol}} = [P,\ \sin(2\Phi),\ \cos(2\Phi)]. \tag{10}$$

As shown in Tab. 11, the proposed decoupled encoding achieves the best Overall score and performs best on four out of five tasks. By representing the polarization degree P and angular structure $(\sin(2\Phi), \cos(2\Phi))$ as separate channels, it preserves orientation cues while avoiding the periodic discontinuity of directly using raw AoLP Φ. The direct S_0+DoLP+AoLP encoding obtains the second-best Overall score and slightly surpasses our decoupled encoding on *reflection recognition*.

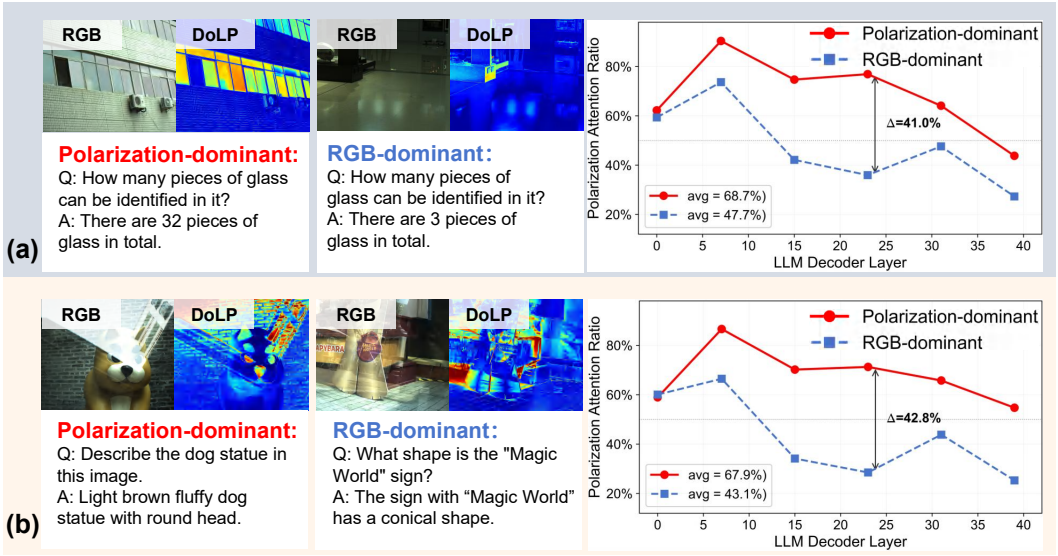

Figure 12: Adaptive attention allocation in PolarVLM. We plot the *Polarization Attention Ratio* across the LLM decoder layers. The results demonstrate that PolarVLM dynamically allocates higher attention to polarimetric tokens (red) for highly ambiguous physical scenes, while relying more on RGB tokens (blue) for simpler queries, proving its capacity for context-aware modality routing without explicit fusion gates.

This is reasonable because *reflection recognition* often depends on coarse specular-region evidence, where the intensity channel S_0 and raw angular response Φ can provide additional cues for identifying reflected content. However, this advantage is task-specific and does not transfer to glass-centric or general scene-description tasks, where the raw AoLP discontinuity and intensity-like bias make the representation less robust overall. In contrast, the DoLP-coupled angular representation suppresses angular responses in weakly polarized regions, which may also attenuate useful directional information. The S_0+Stokes representation performs worst overall, likely because S_0 introduces intensity-like information already correlated with RGB appearance, encouraging the model to underuse more discriminative polarization cues.

Adaptive attention allocation. To gain mechanistic insight into why direct sequence concatenation outperforms explicit fusion modules, we analyze the attention mass from language tokens to visual tokens across the LLM's 40 decoder layers. Specifically, we define the *Polarization Attention Ratio* as the proportion of visual attention weights allocated to polarimetric tokens versus RGB tokens. As illustrated in Fig. 12, the LLM dynamically modulates this ratio based on the inherent optical ambiguity of the scene. For instance, in the *glass counting* task (Fig. 12(a)), a complex facade with multiple indistinguishable glass panels heavily relies on the polarization branch (68.7% average attention). Conversely, a simpler scene with distinct visual boundaries requires significantly less polarimetric grounding (47.7%). A similar content-aware routing emerges under reflection interference (Fig. 12(b)): the model heavily attends to physical priors to decipher severely distorted surfaces, but reverts to RGB dominance when high-level geometric or textual cues (*e.g.*, a printed sign) are sufficient. Layer-wise trajectories further reveal that this modality divergence peaks in the middle decoder layers, where complex semantic reasoning typically occurs. These findings compellingly demonstrate that direct concatenation empowers the LLM to function as a zero-parameter, ambiguity-driven fusion gate, adaptively gathering physical evidence only when standard RGB semantics fall short.

H Polarization Visualizations for General-Purpose VLMs

A potential concern regarding our main evaluation (Tab. 1) is that general-purpose VLMs receive only RGB input, whereas PolarVLM additionally accesses raw polarization measurements. To examine whether this creates an unfair advantage, we augment each baseline with human-readable polarization

visualizations: DoLP is rendered as a grayscale heatmap and AoLP as a cyclic HSV map. For all 1,000 test samples, each model receives the RGB image alongside the labeled DoLP and AoLP visualizations, together with a fixed prompt explaining the physical meaning of each channel. We evaluate three representative baselines—InternVL3-14B (Zhu et al., 2025), Qwen2.5-VL-7B (Bai et al., 2025b), and GPT-4.1 (Achiam et al., 2023)—using the same LLM-as-a-Judge protocol (Wu et al., 2025; Jiang et al., 2026b) as in Sec. 6.

Table 12: General-purpose VLMs with polarization visualizations versus PolarVLM with raw polarization input. Scores follow the 1–10 LLM-as-a-Judge protocol.

Model	Input	Glass Count	Glass Loc.	Glass Desc.	Scene Desc.	Refl. Recog.	Overall
InternVL3-14B	RGB	4.55	5.40	6.55	5.02	2.81	5.01
InternVL3-14B	RGB + Polar Vis.	4.78	5.82	5.89	4.65	3.14	4.85
Qwen2.5-VL-7B	RGB	5.22	5.38	6.65	5.54	3.10	5.42
Qwen2.5-VL-7B	RGB + Polar Vis.	5.40	5.76	6.02	5.12	3.38	5.22
GPT-4.1	RGB	4.81	5.46	5.96	5.66	3.71	5.38
GPT-4.1	RGB + Polar Vis.	4.56	6.13	5.06	4.81	4.04	4.90
PolarVLM	**RGB + Raw Polar.**	**7.57**	**5.51**	**6.87**	**5.85**	**4.09**	**6.08**

As shown in Tab. 12, augmenting the input with polarization visualizations yields mixed per-task effects: glass localization and reflection recognition improve, whereas glass description and scene description degrade. All three models exhibit a net Overall decrease after augmentation. The best augmented configuration (Qwen2.5-VL-7B, Overall 5.22) still trails PolarVLM (6.08) by 0.86 points. These results confirm that rendered polarization heatmaps cannot substitute for a learned physical encoder aligned through structured training, and that the RGB-only comparisons in Tab. 1 do not systematically disadvantage the baselines.

I Inference efficiency

We benchmark each PolarVLM variant against its matched RGB-only counterpart using 100 PolarVQA samples with greedy decoding on a single NVIDIA RTX 5880 Ada GPU (48 GB, BF16 precision, no quantization).

Table 13: Inference efficiency. Deltas in parentheses are relative to the matched RGB-only model.

Model	Input	Vis. Tokens	Latency (ms)	Throughput (sam/s)	Peak Mem. (GB)
LLaVA-1.5-7B	RGB	576	455	2.20	13.9
LLaVA-1.5-7B	Polar	1152	1122 (+147%)	0.89	15.5 (+11%)
LLaVA-1.5-13B	RGB	576	713	1.40	26.1
LLaVA-1.5-13B	Polar	1152	1531 (+115%)	0.65	28.2 (+8%)
Qwen3-VL-8B	RGB	896	929	1.08	16.7
Qwen3-VL-8B	Polar	1793	1159 (+25%)	0.86	18.2 (+9%)
InternVL3.5-14B	RGB	256	1508	0.66	29.4
InternVL3.5-14B	Polar	512	1614 (+7%)	0.62	30.1 (+2%)

The dual-stream design doubles the visual token count. Across the four PolarVLM variants built on LLaVA-1.5 (Liu et al., 2024a), Qwen3-VL (Bai et al., 2025a), and InternVL3.5 (Wang et al., 2025a), latency overhead ranges from 7% (InternVL3.5-14B) to 147% (LLaVA-1.5-7B), while peak memory increases by 2–11%. Memory growth remains moderate because the LLM backbone is unchanged; latency is the primary cost, particularly for LLaVA-1.5 where self-attention over concatenated tokens dominates wall-clock time. Reducing this overhead via token pruning or early-exit strategies merits further investigation.

J Responsible use and asset compliance

PolarVQA is constructed from publicly available polarimetric datasets (Yao et al., 2025; Mei et al., 2022) and does not contain Internet-scraped or private personal data. All source datasets, pretrained models, and codebases are cited, and their respective licenses and terms of use are respected. The

released resources are intended for research on physics-aware visual reasoning. All generated annotations are produced through structured, text-only, rule-verified prompts, validated against ground-truth physical measurements, and subjected to human audit to ensure factual accuracy.