VaseVQA: Multimodal Agent and Benchmark for Ancient Greek Pottery

Jinchao Ge[1*] **Tengfei Cheng**[2*] **Biao Wu**[3*] **Zeyu Zhang**[4*†] **Shiya Huang**[1] **Judith Bishop**[5] **Gillian Shepherd**[5] **Meng Fang**[2] **Ling Chen**[3] **Yang Zhao**[5‡]

[1]University of Adelaide [2]University of Liverpool [3]University of Technology Sydney
[4]The Australian National University [5]La Trobe University

[*]Equal contribution [†]Project lead [‡]Corresponding author: y.zhao2@latrobe.edu.au

Abstract

Understanding cultural heritage artifacts such as ancient Greek pottery requires expert-level reasoning that remains challenging for current MLLMs due to limited domain-specific data. We introduce VaseVQA, a benchmark for ancient Greek pottery, primarily vases, consisting of 31,773 images and 67,614 question–answer pairs across seven expert-defined categories, enabling systematic evaluation of expert-level cultural heritage understanding. Using this dataset, we explore effective training strategies for domain-specific reasoning. While supervised fine-tuning improves adaptation to domain knowledge, it struggles with deeper reasoning tasks. We propose VaseVL, which augments SFT with reinforcement learning using verifiable rewards. Experiments show that VaseVL consistently outperforms supervised baselines, especially on reasoning-intensive questions, highlighting the value of targeted reinforcement learning for cultural heritage visual question answering. Our code and dataset will be released at `https://github.com/AIGeeksGroup/VaseVQA`.

1 Introduction

Cultural heritage artifacts are physical objects that connect us to the past and provide valuable information about art, technology, and society over time. Computational analysis of these artifacts has been a longstanding objective in digital humanities, aiming to augment expert knowledge, facilitate large-scale cataloging, and democratize access to specialized expertise (Castellano and Vessio, 2022). While recent advances in vision-language models have demonstrated remarkable capabilities in general-domain tasks (Li et al., 2023; Liu et al., 2023), their applicability to specialized domains that demand deep expert knowledge remains constrained. Ancient Greek pottery is a good example of why analyzing cultural heritage artifacts is challenging. A single vase can provide different types of information: the clay reveals where it was made, the shape suggests its use and time period, the decoration reflects workshop traditions, and stylistic details help identify specific artists or schools (Smith et al., 2024). Domain experts synthesize these visual cues with extensive contextual knowledge, ranging from regional manufacturing practices to historical timelines and scholarly conventions, to answer complex questions about where the objects come from, when they were made, and who made them. Automating this analysis requires models that can both recognize fine visual details and reason using specialized domain knowledge—capabilities that go beyond what current general-purpose MLLMs can provide.

However, most existing vision-language models lack the domain-specific knowledge required to answer these fundamental questions reliably. Such knowledge is inherently scarce in the real world, as it is typically confined to expert literature, museum archives, and specialized scholarship. As a result, current VLMs are underexposed to this type of information during pre-training, which limits their ability to reason about cultural heritage artifacts beyond surface-level visual cues.

To address this gap, we curate and construct a dedicated dataset that systematically captures expert knowledge for ancient Greek pottery, providing a foundation for training and evaluating models on these core archaeological questions. To facilitate systematic evaluation, we introduce VaseVQA, a comprehensive benchmark comprising 31,773 images (featuring an 11,693-image single-view subset) of primarily ancient Greek vases, together with 67,614 visual question–answer pairs. The benchmark encompasses seven distinct question types and incorporates type-specific evaluation metrics, enabling rigorous assessment of both lexical precision and semantic alignment.

With the dataset in place, a natural question is how existing vision–language models can be

Findings of the Association for Computational Linguistics: EACL 2026, pages 1154–1167

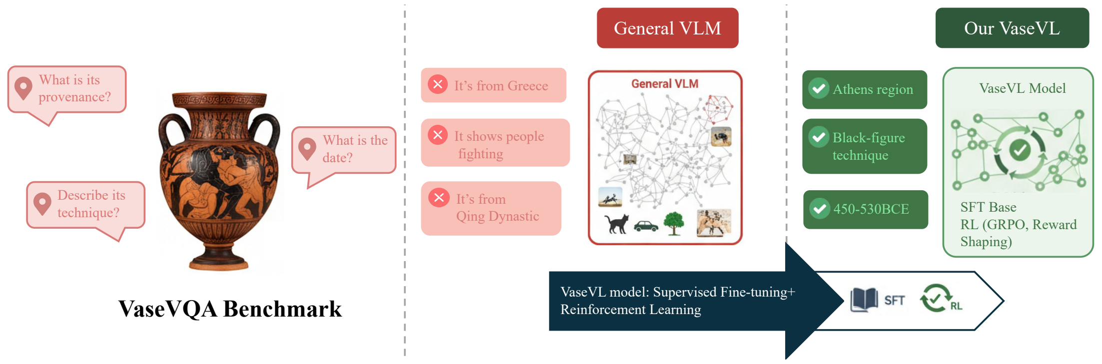

Figure 1: Bridging the Semantic Gap: Limitations of Existing Vision-Language Models on Ancient Greek Vase Understanding and the Proposed VaseVL Framework

adapted to reason over such expert knowledge. Since the annotations in VaseVQA lie largely outside the pre-training distribution of current MLLMs, standard supervised fine-tuning can partially bridge this gap by improving factual recall. However, our empirical analysis shows that exposure to expert data alone is insufficient for robust reasoning over attributes that are not directly observable, such as provenance, dating, or attribution. This observation motivates the use of additional training signals beyond conventional supervision to better align models with expert-level reasoning requirements. We define a seven-category question taxonomy—*Fabric*, *Technique*, *Shape*, *Provenance*, *Attribution*, *Date*, and *Decoration*—and use it to identify type-specific weaknesses in post-SFT models. Based on this taxonomy, we design a rule-based reward to guide reinforcement learning, aiming to improve performance on more challenging questions by strengthening the model's reasoning ability.

Experimental results show that general-purpose MLLMs still face a clear domain gap in cultural heritage understanding. Although these models achieve strong baseline performance on general tasks, they struggle with expert-level questions in zero-shot settings, largely due to the scarcity of domain-specific knowledge and the lack of specialized training data for artifacts such as Ancient Greek pottery. SFT substantially improves factual recall, achieving near-ceiling performance on visually explicit categories (e.g., *Fabric* and *Technique*), but remains unstable on reasoning-intensive questions, indicating its limited ability to model deeper contextual relationships. In contrast, VaseVL addresses these limitations by introducing verifiable rewards that explicitly supervise reasoning outcomes, consistently outperforming the SFT baseline on challenging question types while also improving compositional robustness.

Our main contributions are as follows:

- We introduce VaseVQA, a benchmark comprising 31,773 images and 67,614 question–answer pairs. This work fills a critical gap in the field by providing a large-scale, expert-annotated dataset for Ancient Greek vases and porcelain, covering seven expert-defined question types with type-specific evaluation metrics.

- We propose VaseVL, a model developed through a two-stage training process with taxonomy-aware reward shaping, designed to better leverage the expert annotations provided by VaseVQA beyond surface-level visual features.

- Experimental results show that while SFT plateaus in reasoning on this dataset, our targeted RL approach consistently improves performance on challenging question types, highlighting the value of high-quality domain-specific data for expert-level cultural heritage analysis.

2 Related Work

Multimodal Large Language Models (MLLMs). Pretrained vision–language models learn joint representations from large-scale multimodal corpora and have advanced a wide range of tasks, including image–text retrieval, VQA, grounding, and dialogue (Li et al., 2019; Chen et al., 2020; Zeng et al., 2021; Li et al., 2021; Song et al., 2025a,c,b; Huang et al., 2025; Liu et al., 2025). Building on this, Visual Instruction Tuning (VIT) further enhances MLLMs' ability to understand and execute complex multimodal instructions, as exemplified by LLaVA (Liu et al., 2024), MiniGPT-4 (Zhu et al.,

Datasets	Images	Questions	Question Type	Image Type	Task Focus	Venue	OE/MC
DAQUAR(Malinowski and Fritz, 2014)	1,449	12,468	4	Natural	VQA	NIPS 2014	OE
COCO-QA(Ren et al., 2015a)	123,287	117,684	4	Natural	VQA	-	OE
VAQ V1.0(Agrawal et al., 2017)	204k	614K	-	Natural	VQA	ICCV 2015	OE
VQA V2.0(Goyal et al., 2017)	204k	1.1M	-	Natural	VQA	CVPR 2017	Both
CVR(Zellers et al., 2019)	110k	290K		Natural	VR	CVPR 2019	MC
GQA(Hudson and Manning, 2019)	113,018	22,669,678	-	Natural	VR	CVPR 2019	OE
RAVEN(Zhang et al., 2019)	1,120,000	70,000	4	Natural	VR	CVPR 2019	MC
NLVR(Suhr et al., 2017)	387,426	31,418	-	Synthetic	VR	ACL 2019	OE
OK-VQA(Marino et al., 2019)	14,031	14,055	VQA	Natural	VQA	CVPR 2019	OE
VizWiz(Gurari et al., 2018)	-	31,173	-	Natural	-	CVPR 2018	OE
KVQA(Shah et al., 2019)	24K			Natural	-	AAAI 2019	OE
CLEVR(Johnson et al., 2017)	100,000	999,968	90	Natural	VR	CVPR 2017	OE
FM-IQA(Gao et al., 2015)	158,392	316,193	–	Natural	-	CVPR 2017	OE
NLVR2(Suhr and Artzi, 2019)	107,292	29,680	-	Synthetic	VR	ACL 2019	OE
TextVQA(Singh et al., 2019)	28,408	45,336	-	Natural	VQA	CVPR 2019	OE
FVQA(Wang et al., 2017)	2190	5826	12	Natural	VR	CVPR 2019	OE
VISUAL GENOME(Krishna et al., 2017)	108,000	145,322	7	Natural	VR	-	OE
VQA-CP(Agrawal et al., 2018)				Natural	VQA	CVPR 2018	
Visual Madlibs(Yu et al., 2015)	10,738	360,001	12	Natural	VR	-	-
SHAPES(Andreas et al., 2015)	15,616	244	–	Synthetic	VR	-	Binary
KB-VQA(Wang et al., 2015)	700	2402	23	Natural	VQA	IJCAI 17	OE
ICQA(Hosseinabad et al., 2021)	42,021	260,840	-	Synthetic	VQA	-	OE
DVQA(Kafle et al., 2018)	3,000,000	3,487,194	3	-	VQA	-	OE
PathVQA(He et al., 2020)	4,998	32,795	7	-	VQA	-	MC
Visual7w(Zhu et al., 2016a)	47,300	327,939	7	Natural	VQA	CVPR 16	MC
KRVQA(Cao et al., 2021)	32,910	157,201	6	Natural	VR	-	MC
VaseVQA	11,693	67,614	7	Natural	VQA	-	OE

Table 1: A Main characteristics of major VQA and Visual Reasoning datasets.

2023), Gemini 1.5 (Google, 2024), and Qwen2.5-VL-Instruct (Team, 2025). As shown in Table 7 in the appendix, current MLLMs are typically evaluated across a broad suite of vision tasks, validating their general perception and cross-modal alignment capabilities. However, such generic evaluations remain insufficient for highly specialized scenarios such as cultural-heritage analysis, which require integrating fine-grained visual cues with historically and archaeologically grounded knowledge to support expert-level reasoning and judgment.

VQA Benchmarks. Foundational VQA datasets, including VQA (Antol et al., 2015), COCO-QA (Ren et al., 2015b), and Visual7W (Zhu et al., 2016b), enabled rapid progress but primarily cover generic objects and scenes. In cultural-heritage domains, publicly available resources remain scarce. RePAIR (Tsesmelis et al., 2024) targets oracle bones, and HUST-OBS (Wang et al., 2024) focuses on fragment reconstruction, leaving a gap for classical artifacts such as ancient Greek vases. Table 1 summarizes the key characteristics of major VQA and visual reasoning datasets. Our VaseVQA fills this gap with a VQA-centric benchmark designed to probe factual recall (e.g., *Fabric*, *Technique*) and expert-level reasoning (*Attribution*, *Decoration*, *Date*, *Provenance*, *Shape*) under a type-aware evaluation protocol.

3 Data Collection

VaseVQA was constructed with a focus on representing Ancient Greek culture, ensuring both the visual and textual components accurately reflect the region's rich cultural heritage. The Table 1 presents major VQA and Visual Reasoning datasets, compared with the VaseVQA. Table 3 shows the division of the VaseVQA dataset into training and test sets, with each image having 7 questions. In addition, Table 2 illustrates examples within the dataset, demonstrating the questions crafted for every sample.

3.1 Image Collection

The images in this dataset were collected through collaborations with Ancient Greek archaeological institutions, museums, and cultural heritage centers (Classical Art Research Centre, University of Oxford, 2025). We focused on gathering images of classical funerary vases that are commonly found in Ancient Greek archaeological sites, ensuring a diverse representation of artifacts across different cultural groups. The images include both complete objects and fragments, as well as images of the vases in their original burial contexts. This collection was designed to capture intricate details of

VaseVQA example:	
	
Question Answer	What is the fabric of the vase? The fabric of the vase is ATHENIAN.
Question Answer	What is the technique of the vase? The technique of the vase is RED-FIGURE.
Question Answer	What is the shape name of the vase? The shape name of the vase is CUP B.
Question Answer	What is the provenance of the vase? The provenance of the vase is not available.
Question Answer	What is the date of the vase? The date of the vase is -450 to -400.
Question Answer	What is the attribution of the vase? The vase is attributed to CODRUS P by BURN | CODRUS P by UNKNOWN.
Question Answer	What is the decoration of the vase? The decoration of the vase is A,B: THEATRICAL, DRAPED SATYRS, WITH STORK, BOX AND SANDAL, ARYBALLOI, OINOCHOE, LYRE AND STAFFS, ONE CONFRONTING DRAPED YOUTH | I: AMAZON ON HORSEBACK.

Table 2: **Example of a VaseVQA Dataset**: Eight attribute-specific questions (fabric, technique, shape, provenance, date, attribution, decoration) paired with visual input, presented in a conversational Q&A format to analyze an Athenian red-figure cup (450–400 BCE) attributed to the Codrus Painter.

materials, craftsmanship, and regional variations on the Table 2.

3.2 Text Collection

The textual data for the dataset was derived from several key sources, including academic papers, archaeological reports, and expert annotations provided by Ancient Greek historians and cultural heritage experts. The texts are descriptions of the artifacts, detailing their material composition (such as red pottery or glazed ceramics), motifs (such as human, animal, or abstract designs), and the archaeological context (such as burial sites or ceremonial uses). These descriptions were translated and structured to align with the images and make the data accessible for vision-language tasks.

3.3 Annotation

The dataset was labeled by a team of archaeologists and cultural heritage experts, who annotated the images with several key attributes. These include the material (e.g., red pottery, glazed ceramics), pattern type (e.g., human figures, animal motifs, abstract symbols), excavation layer, radiocarbon dating estimates, manufacturing techniques (e.g., hand-built, wheel-thrown, firing temperature), and the contextual use of the object (e.g., funerary or ceremonial). The experts also identified restoration marks, if applicable. The labeling process ensures a high level of detail and accuracy in reflecting the cultural and historical context of each artifact.

4 Methods

We start from a supervised fine-tuned (SFT) model trained on the ancient Greek vase dataset. While the model demonstrates basic visual question-answering capability, it exhibits systematic weaknesses in reasoning and compositional understanding. We apply reinforcement learning to address these limitations, enhancing reasoning ability while preserving foundational knowledge.

We denote the post-SFT model as the reference policy and initialize a trainable policy from it. The objective is to achieve expert-level accuracy and compositional robustness. To ensure stable training across diverse question types, we employ Group Relative Policy Optimization (GRPO). For each image-question pair, we sample multiple candidate answers and compute type-conditioned rewards. GRPO normalizes these rewards per prompt by subtracting the average, yielding relative advantages that are robust to cross-type scale variation. We then optimize a clipped PPO-style objective

Split	Fabric	Technique	Shape	Provenance	Date	Attribution	Decoration
Question Type Distribution (%)							
Train	17.3	17.3	17.3	6.7	17.3	7.0	17.1
Test	17.2	17.2	17.2	6.9	17.2	7.1	17.1
Average Answer Length (words)							
Avg.	10.0	11.0	13.0	16.6	12.0	20.9	28.3
Total	67,614 Question–Answer Pairs						

Table 3: Dataset composition of VaseVQA. The dataset is split into training and test sets with a 4:1 ratio. Question-type distributions are nearly identical across splits. Attribution and Decoration exhibit longer average answers, reflecting their more open-ended response requirements.

with KL regularization to prevent drift from the reference policy (Figure 2).

To make GRPO effective, we design a reward engineering framework tailored to the shortcomings of the SFT model. The reward consists of two complementary metrics: (1) a keyword-based score measuring lexical overlap, prioritizing factual correctness, and (2) a semantic similarity score based on normalized cosine similarity of sentence embeddings. Recognizing that different tasks weigh precision versus semantics differently, we combine them with adaptive, type-conditioned weights. Furthermore, for error-prone categories, we amplify the reward signal with an additional weighting factor, focusing learning on areas where the SFT model is weakest.

4.1 Reward Design

To effectively guide model optimization, we propose a comprehensive reward function that directly addresses the limitations of the SFT model and flexibly adapts to different categories of questions. The reward integrates both lexical and semantic perspectives, ensuring that generated answers $\hat{a}$ are evaluated not only for factual precision but also for meaning preservation with respect to the ground-truth reference a^*.

Keyword-based Score ($s_{\mathbf{kw}}$): Lexical accuracy is particularly important for factual and knowledge-intensive queries. To capture this, we define a keyword-based score that measures the overlap of essential terms between the model output and the reference:

$$s_{\text{kw}} = \frac{|K(\hat{a}) \cap K(a^*)|}{|K(\hat{a}) \cup K(a^*)|}, \tag{1}$$

where $K(\cdot)$ extracts the keyword set from a given text. This formulation is analogous to a Jaccard similarity over keywords, prioritizing factual correctness and penalizing omissions or hallucinated entities.

Semantic Similarity Score ($s_{\mathbf{sem}}$): While keyword overlap captures precision, it may fail to reflect semantic equivalence when paraphrases are used. To complement this, we compute semantic similarity using the `all-MiniLM-L6-v2` model from the Sentence-Transformers library. Each text string is embedded into a high-dimensional vector space, and the similarity is defined as the normalized cosine similarity:

$$s_{\text{sem}} = \frac{\cos(f(\hat{a}), f(a^*)) + 1}{2}, \tag{2}$$

where $f(\cdot)$ denotes the embedding function from (Wang et al., 2020). This design ensures that semantically faithful answers, even if phrased differently, receive appropriate credit.

4.2 Reward Shaping

A critical insight is that the importance of lexical accuracy versus semantic equivalence is not uniform across tasks. For example, in factual questions such as *"What is the capital of Canada?"*, exact keyword matching ("Ottawa") is indispensable. In contrast, for descriptive or open-ended queries like *"Explain the impact of climate change"*, semantic similarity plays a more dominant role, as diverse phrasings can still convey correct meaning.

To incorporate this adaptivity, we define the reward as a weighted combination:

$$\tilde{R}(q) = \beta(q)s_{\text{kw}} + (1 - \beta(q))s_{\text{sem}}, \tag{3}$$

where the weight $\beta(q)$ is conditioned on the question type q. This weight allows the reward to adapt its emphasis to the requirements of different tasks.

Furthermore, to emphasize learning in areas where the SFT model is weakest, we apply an additional scaling factor:

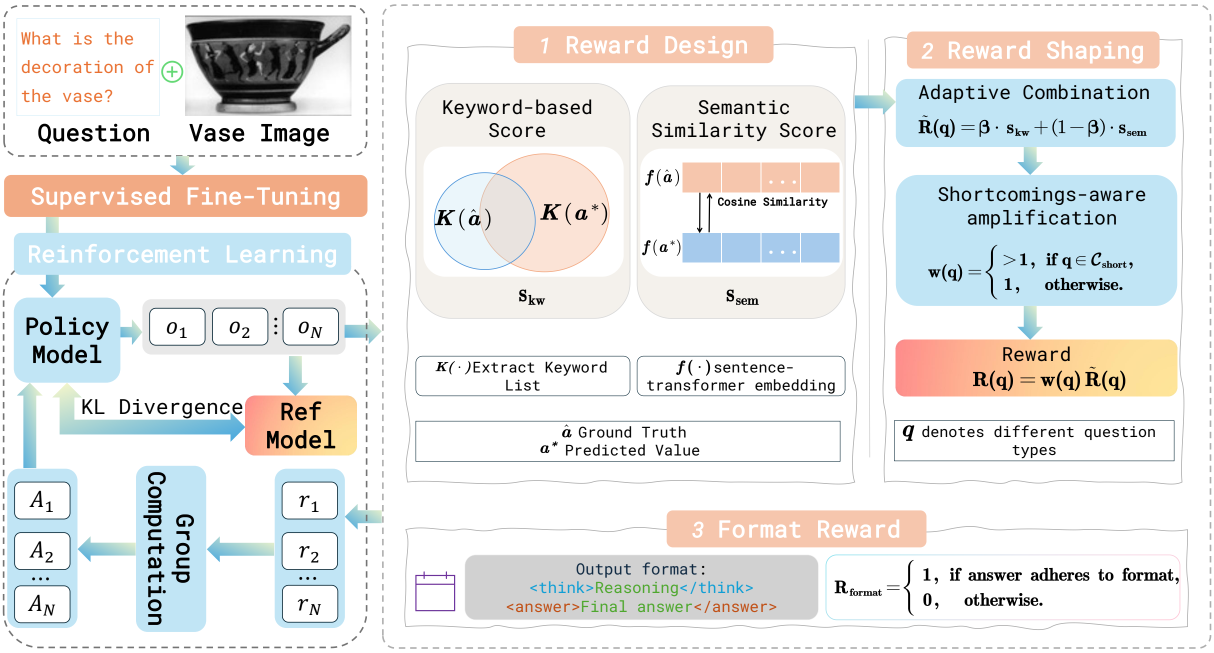

Figure 2: Overall framework of VaseVL. The proposed pipeline integrates SFT with RL under the GRPO paradigm. Given a vase image x, a question q, and the reference answer a^*, the model refines its reasoning ability by balancing lexical and semantic rewards while constraining policy drift from π_{ref}.

$$R(q) = w(q) \cdot \tilde{R}(q), \tag{4}$$

with

$$w(q) = \begin{cases} > 1, & \text{if } q \in \mathcal{C}_{\text{short}}; \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

Here, $\mathcal{C}_{\text{short}}$ denotes a predefined set of question types where the SFT model is known to underperform (e.g., numerical reasoning, multi-step factual queries). By amplifying the reward signal in these regions, the training process allocates greater attention to challenging categories, accelerating targeted improvements without neglecting overall performance.

5 Experiments

In this section, we conduct a series of experiments to validate the effectiveness of our proposed model, VaseVL. We evaluate our model on the newly introduced VaseVQA benchmark and compare its performance against various strong baselines, including general-purpose Multimodal Large Language Models (MLLMs) and a fine-tuned version of the model without reinforcement learning. Our evaluation is structured around the taxonomy of seven distinct question types to provide a granular analysis of the model's capabilities in expert-level reasoning.

5.1 Implementation details

We propose a two-stage training paradigm. In the first stage, we perform full-parameter supervised fine-tuning (SFT) on the MLLMs, and in the second stage, we conduct task-specific reinforcement learning (RL) training. Concretely, we initialize from a general-purpose MLLM and instruction-tune on $\mathcal{D} = \{(x_i, q_i, a_i^*)\}_{i=1}^{N}$.

The SFT model π_{ref} is used both as a stable reference for RL and as a probe to evaluate per-type performance over $\mathcal{T}$ (e.g., *Fabric*, *Technique*), from which we select a shortcoming subset $\mathcal{C}_{\text{short}} \subseteq \mathcal{T}$ for targeted improvement. For SFT, we adopt a per-device batch size of 1 with $8\times$ gradient accumulation, a cosine learning rate schedule with initial value 1×10^{-4}, one epoch of training, a warmup ratio of 0.1, and enable `bf16`. After SFT, we perform RL focused on $\mathcal{C}_{\text{short}}$ with a GRPO-style setup: 8 rollouts per prompt, temperature 0.9, one iteration per batch with KL penalty coefficient 0.04, training for 2 epochs at learning rate 1×10^{-6}.

5.2 Evaluation Metrics

We use a task-specific evaluation protocol, applying the most appropriate metric to each question type. We report accuracy for factual question types, with task-specific computation protocols.

ANLS-based Accuracy For short-answer factual questions—*Fabric*, *Technique*, *Shape*, *Provenance*,

Model	Param.	Fabric	Technique	Shape	Provenance	Date	Attribution	Decoration	
		Accuracy						BLEU@1	Overall
LLaVA (Liu et al., 2024)	7B	11.56	0	44.60	0	0	28.41	6.28	14.10
Vicuna (Zheng et al., 2023)	7B	0	0	0.24	0	0	0	1.14	0.04
MiniCPM (Yao et al., 2024)	8B	0.29	0.03	0	0.97	0	0.14	1.15	0.24
InternVL2 (Chen et al., 2024)	1B	10.50	0.08	3.97	11.52	0	14.10	3.41	6.70
InternVL2 (Chen et al., 2024)	2B	0	0.60	6.03	8.62	0.65	3.99	1.94	3.31
InternVL2 (Chen et al., 2024)	8B	1.69	4.00	7.66	21.24	0	2.85	2.00	6.24
Qwen-2.5-VL (Team, 2025)	3B	0.29	0.02	0.14	0.00	14.86	0.00	2.29	2.55
Qwen-2.5-VL (Team, 2025)	7B	0.00	0.00	0.05	0.00	17.95	0.00	1.91	3.00
Gemini-2.0-Flash	–	9.37	90.11	26.22	12.42	17.41	2.92	6.79	26.41
GPT-4o-mini	–	80.21	86.08	76.95	68.84	**47.22**	39.51	6.53	66.47
VaseVL (Ours)	3B	**99.95**	**95.93**	**83.99**	**73.67**	39.87	**60.83**	**9.82**	**75.71**

Table 4: Performance comparison on the VaseVQA benchmark.

and *Attribution*—we compute accuracy using Average Normalized Levenshtein Similarity (ANLS), a ST-VQA–inspired soft metric robust to minor character-level and OCR-like errors (Biten et al., 2019).

Date Accuracy For *Date* questions, we parse numerical years and define accuracy as the maximum over three components: partial credit for correct date-range formatting, a proximity-based score within a tolerance margin, and a primary score based on the Intersection over Union (IoU) of year ranges.

BLEU@1 Score For the descriptive *Decoration* questions, the target output consists of a list of specific visual attributes and keywords, with an emphasis on lexical precision. Accordingly, we adopt BLEU@1 as the evaluation metric, measuring unigram precision between the generated outputs and the ground truth. This metric effectively captures key descriptive terms without overly penalizing stylistic variations, and serves as a proxy for the model's compositional understanding.

5.3 Main Results

Effect of Model Scale The results in Table 4 indicate that simply increasing model size does not lead to performance gains on VaseVQA. Larger general-purpose MLLMs, such as LLaVA 7B (Liu et al., 2024), Vicuna 7B (Zheng et al., 2023), and MiniCPM 8B (Yao et al., 2024), fail to outperform smaller models and in some expert-level tasks even achieve scores close to zero. This outcome is primarily due to the lack of cultural-heritage knowledge in their pretraining data: ancient Greek pottery styles, shapes, and historical contexts are almost absent from web-scale corpora. In contrast, VaseVL, with only 3B parameters, integrates supervised fine-tuning to supplement domain knowledge and diagnosis-guided reinforcement learning to activate reasoning. This combination allows our model to surpass all larger baselines. The result underscores that domain-specific data and training strategies are more critical than raw model scale, enabling smaller, efficient models to achieve expert-level performance.

Task-Type Variability Low-level tasks such as *Fabric* and *Technique* fail in zero-shot settings due to missing domain-specific textures, but reach near-perfect accuracy after supervised fine-tuning, indicating a reliance on visual feature coverage rather than explicit reasoning. Mid-level tasks including *Shape* and *Provenance* exhibit limited zero-shot capability, likely inherited from analogous patterns in pretraining data, yet achieving expert-level performance requires domain alignment through supervised fine-tuning and reinforcement learning. High-level tasks—*Date*, *Attribution*, and *Decoration*—expose fundamental limitations of current VLMs: zero-shot models largely collapse, and although VaseVL attains substantial improvements, with 39.87% on *Date*, 60.83% on *Attribution*, and 9.82 BLEU@1 on *Decoration*, a significant gap to easier tasks remains. These tasks demand historical reasoning, multi-evidence integration, and compositional language generation beyond pure visual recognition. Overall, low-level perception adapts rapidly with data exposure, mid-level reasoning benefits from domain alignment, while high-level historical reasoning remains a core bottleneck in current vision–language modeling.

SFT	RL	Fabric	Technique	Shape	Provenance	Date	Attribution	Decoration		
		Accuracy						BLEU@1	Prometheus Score	Overall
-	-	0.29	0.02	0.14	0.00	14.86	0.00	2.29	0.99	2.55
-	✓	13.33	19.95	14.82	5.27	3.58	11.50	4.82	2.91	11.41
✓	-	**99.96**	94.99	83.98	71.67	37.96	56.96	2.57	10.31	74.25
✓	✓	99.95	**95.93**	**83.99**	**73.67**	**39.87**	**60.83**	**9.82**	**14.71**	**75.71**

Table 5: Ablation study results. The final row reports the performance of our proposed VaseVL. Prometheus Score is the LLM-as-a-judge score produced by *Prometheus-v2.0-7B*.

VaseVQA-mini	Fabric	Technique	Shape	Provenance	Date	Attribution	Decoration	
	Accuracy						BLEU@1	Overall
w/o Keyword (s_{kw})	99.95	94.62	83.43	72.27	38.13	57.14	5.93	74.25
w/o Semantic (s_{sem})	99.95	94.83	83.74	72.86	37.53	56.94	3.24	74.30
w/o Amplification ($w(q)$)	99.96	**94.97**	**84.02**	73.05	38.04	57.23	4.13	74.54
VaseVL	**99.96**	94.93	83.98	**73.24**	**38.55**	**57.79**	**6.21**	**74.74**

Table 6: Performance comparison on the VaseVQA benchmark.

5.4 Ablation Study

Effect of Training Stages Table 5 compares the contributions of Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL). The zero-shot baseline's low overall score (2.55) highlights the domain gap in general-purpose models. RL without prior SFT yields only marginal gains in reasoning (e.g., 11.50% on *Attribute*), lacking sufficient grounding. Conversely, SFT-only excels in factual recall (*Fabric*: 99.96%, *Technique*: 94.99%) but struggles with complex generation, evidenced by low scores on *Date* (37.96%) and *Decoration* (2.57 BLEU@1). VaseVL integrates both to achieve superior performance (75.71% overall), retaining factual accuracy while significantly boosting reasoning (e.g., *Decoration* improves to 9.82 BLEU@1). Additional evaluation using Prometheus-7B-v2.0 (Kim et al., 2024) as an LLM-judge further corroborates the efficacy of this two-stage strategy in enhancing descriptive quality.

Effect of Reward Components We analyze reward efficacy on *VaseVQA-mini* (a stratified 20% subset) in Table 6. Ablating the Keyword Score (s_{kw}) impairs strict classification (*Technique* drops to 94.62%, *Shape* to 83.43%), confirming the need for lexical precision. Removing the Semantic Score (s_{sem}) degrades open-ended generation, with *Decoration* plummeting from 6.21 to 3.24 BLEU-1, indicating the importance of embedding-based guidance. Finally, excluding the Shortcomings-aware Amplification ($w(q)$) hinders reasoning on difficult tasks like *Date* (38.04%) and *Attribution* (57.23%). The full VaseVL framework achieves the highest overall score (74.74%), validating the synergistic value of these reward signals.

6 Conclusion

We introduced VaseVL, a model trained with a two-stage framework and reward shaping for cultural heritage analysis, together with VaseVQA, a comprehensive benchmark spanning seven question types. By aligning optimization with task-specific evaluation metrics and amplifying rewards for complex reasoning, VaseVL moves beyond surface-level pattern recognition. Specifically, GRPO with KL regularization preserves factual accuracy while enabling expert-level reasoning. The release of VaseVQA establishes a reproducible standard for the field. Our findings on ancient Greek pottery indicate that reinforcement learning with targeted reward shaping is a promising direction for adapting MLLMs to other domains that require expert-level reasoning. Future work will validate this approach on a broader range of cultural heritage datasets.

7 Social Impact

VaseVL, a foundational MLLM for Ancient Greek pottery, preserves and interprets this vital cultural heritage by integrating visual and textual analysis of styles, shapes, decorations, and inscriptions. It aids archaeologists in accurate classification and digital documentation, enhances access to global collections, and supports forgery detection and cultural preservation. Museums and educators can also use VaseVL to create engaging learning experiences. Ultimately, VaseVL bridges AI and archaeology, safeguarding the historical legacy of Ancient Greek pottery for future generations.

8 Limitations

Although VaseVQA provides a large-scale benchmark with carefully designed training and test splits, data bias and coverage remain important limitations. The dataset focuses on ancient Greek pottery, and while it captures a wide range of fabrics, techniques, shapes, and decorative styles, it cannot fully represent the diversity of cultural heritage artifacts across different regions, periods, and materials. In addition, question–answer distributions may reflect annotation preferences and existing scholarly conventions, which could inadvertently bias model training. As a result, models optimized on VaseVQA may overfit to dataset-specific patterns rather than generalize to real-world archaeological or art-historical contexts. Future work should address these issues by incorporating more heterogeneous artifacts, broader cultural traditions, and multiple sources of expert annotation to improve robustness and reduce bias.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.

Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. 2017. C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset. *arXiv preprint arXiv:1704.08243*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2015. Deep compositional question answering with neural module networks. corr abs/1511.02799 (2015). *arXiv preprint arXiv:1511.02799*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. *2015 IEEE International Conference on Computer Vision (ICCV), IEEE*, pages 2425–2433.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. 2014. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, pages 2011–2018.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer.

Qingxing Cao, Bailin Li, Xiaodan Liang, Keze Wang, and Liang Lin. 2021. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Transactions on Neural Networks and Learning Systems*.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*.

Giovanna Castellano and Gennaro Vessio. 2022. A deep learning approach to clustering visual arts. *International Journal of Computer Vision*, 130(11):2590–2605.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *CVPR*, pages 3606–3613.

Classical Art Research Centre, University of Oxford. 2025. Beazley archive pottery database (bapd). https://www.carc.ox.ac.uk/pottery/. Accessed 2025-09-02.

Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pages 215–223. JMLR Workshop and Conference Proceedings.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, pages 178–178. IEEE.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. *arXiv preprint arXiv:1505.05612*.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE.

Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, and 1 others. 2013. Challenges in representation learning: A report on three machine learning contests. In *ICONIP*, pages 117–124. Springer.

Gemini Team Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *JSTARS*, 12(7):2217–2226.

Sayedshayan Hashemi Hosseinabad, Mehran Safayani, and Abdolreza Mirzaei. 2021. Multiple answers to a question: a new approach for visual question answering. *The Visual Computer*, 37(1):119–131.

Ting Huang, Zeyu Zhang, and Hao Tang. 2025. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. *arXiv preprint arXiv:2507.23478*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *NeurIPS*, 33:2611–2624.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, and 1 others. 2022. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *arXiv preprint arXiv:2204.08790*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Qingxiang Liu, Ting Huang, Zeyu Zhang, and Hao Tang. 2025. Nav-r1: Reasoning and navigation in embodied scenes. *arXiv preprint arXiv:2509.10884*.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27:1682–1690.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. Rareact: A video dataset of unusual interactions. *arXiv preprint arXiv:2008.01018*.

Anand Mishra, Karteek Alahari, and CV Jawahar. 2012. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pages 891–898.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015a. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28:2953–2961.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015b. Exploring models and data for image question answering. *arXiv:1505.02074*.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8876–8884.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Tyler Jo Smith, Ethan Gruber, and Nicholas A Harokopos. 2024. 22 kerameikos. org and digital accessibility for ancient greek vases. *Technology, Crafting and Artisanal Networks in the Greek and Roman World: Interdisciplinary Approaches to the Study of Ceramics*, page 255.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642.

Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Chenxi Wang, Guangxian Ouyang, and 1 others. 2025a. Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models. *arXiv preprint arXiv:2505.15406*.

Zirui Song, Guangxian Ouyang, Mingzhe Li, Yuheng Ji, Chenxi Wang, Zixiang Xu, Zeyu Zhang, Xiaoqing Zhang, Qian Jiang, Zhenhao Chen, and 1 others. 2025b. Maniplvm-r1: Reinforcement learning for reasoning in embodied manipulation with large vision-language models. *arXiv preprint arXiv:2505.16517*.

Zirui Song, Jingpu Yang, Yuan Huang, Jonathan Tonglet, Zeyu Zhang, Tao Cheng, Meng Fang, Iryna Gurevych, and Xiuying Chen. 2025c. Geolocation with real human gameplay data: A large-scale dataset and human-like reasoning framework. *arXiv preprint arXiv:2502.13759*.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. 2011. The german traffic sign recognition benchmark: a multi-class classification competition. In *IJCNN*, pages 1453–1460. IEEE.

Alane Suhr and Yoav Artzi. 2019. Nlvr2 visual bias analysis. *arXiv preprint arXiv:1909.10411*.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223.

Qwen Team. 2025. Qwen2.5-vl.

Theodore Tsesmelis, Luca Palmieri, Marina Khoroshiltseva, Adeela Islam, Gur Elkin, Ofir Itzhak Shahar, Gianluca Scarpellini, Stefano Fiorini, Yaniv Ohayon, Nadav Alali, Sinem Aslan, Pietro Morerio, Sebastiano Vascon, Elena Gravina, Maria Cristina Napolitano, Giuseppe Scarpati, Gabriel Zuchtriegel, Alexandra Spühler, Michel E. Fuchs, and 4 others. 2024. Re-assembling the past: The repair dataset and benchmark for real world 2d and 3d puzzle solving. *arXiv:2410.24010*.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

Pengjie Wang, Kaile Zhang, Xinyu Wang, Shengwei Han, Yongge Liu, Jinpeng Wan, Haisu Guan, Kuang Zhebin, Lianwen Jin, Xiang Bai, and Yuliang Liu. 2024. An open dataset for oracle bone script recognition and decipherment. *arXiv:2401.15365*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. 2015. Visual madlibs: Fill in the blank image generation and question answering. *arXiv preprint arXiv:1506.00278*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5317–5327.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641.

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

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016a. Visual7W: Grounded Question Answering in Images. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016b. Visual7w: Grounded question answering in images. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE*, page 4995–5004.

Appendix

Current MLLMs are typically evaluated under a broad suite of visual tasks, including image classification, image–text retrieval, object detection, semantic segmentation, and general-purpose VQA, as summarized in Table 7.

Task	Dataset	Year	Classes	Training	Testing	Evaluation Metric
Image Classification	MNIST (LeCun et al., 1998) [link]	1998	10	60,000	10,000	Accuracy
	Caltech-101 (Fei-Fei et al., 2004) [link]	2004	102	3,060	6,085	Mean Per Class
	PASCAL VOC 2007 Classification (Everingham et al., 2010) [link]	2007	20	5,011	4,952	11-point mAP
	Oxford 102 Folwers (Nilsback and Zisserman, 2008) [link]	2008	102	2,040	6,149	Mean Per Class
	ImageNet-1k (Deng et al., 2009) [link]	2009	1000	1,281,167	50,000	Accuracy
	SUN397 (Xiao et al., 2010) [link]	2010	397	19,850	19,850	Accuracy
	SVHN (Netzer et al., 2011) [link]	2011	10	73,257	26,032	Accuracy
	STL-10 (Coates et al., 2011) [link]	2011	10	1,000	8,000	Accuracy
	GTSRB (Stallkamp et al., 2011) [link]	2011	43	26,640	12,630	Accuracy
	KITTI Distance (Geiger et al., 2012) [link]	2012	4	6,770	711	Accuracy
	IIIT5k (Mishra et al., 2012) [link]	2012	36	2,000	3,000	Accuracy
	Oxford-IIIT PETS (Parkhi et al., 2012) [link]	2012	37	3,680	3,669	Mean Per Class
	FGVC Aircraft (Maji et al., 2013) [link]	2013	100	6,667	3,333	Mean Per Class
	Facial Emotion Recognition 2013 (Goodfellow et al., 2013) [link]	2013	8	32,140	3,574	Accuracy
	Rendered SST2 (Socher et al., 2013) [link]	2013	2	7,792	1,821	Accuracy
	Describable Textures (DTD) (Cimpoi et al., 2014) [link]	2014	47	3,760	1,880	Accuracy
	Food-101 (Bossard et al., 2014) [link]	2014	102	75,750	25,250	Accuracy
	Birdsnap (Berg et al., 2014) [link]	2014	500	42,283	2,149	Accuracy
	RESISC45 (Cheng et al., 2017) [link]	2017	45	3,150	25,200	Accuracy
	CLEVR Counts (Johnson et al., 2017) [link]	2017	8	2,000	500	Accuracy
	PatchCamelyon (Veeling et al., 2018) [link]	2018	2	294,912	32,768	Accuracy
	EuroSAT (Helber et al., 2019) [link]	2019	10	10,000	5,000	Accuracy
	Hateful Memes (Kiela et al., 2020) [link]	2020	2	8,500	500	ROC AUC
	Country211 (Radford et al., 2021) [link]	2021	211	43,200	21,100	Accuracy
Image-Text Retrieval	Flickr30k (Young et al., 2014) [link]	2014	-	31,783	-	Recall
	COCO Caption (Chen et al., 2015) [link]	2015	-	82,783	5,000	Recall
Action Recognition	UCF101 (Soomro et al., 2012) [link]	2012	101	9,537	1,794	Accuracy
	Kinetics700 (Carreira et al., 2019) [link]	2019	700	494,801	31,669	Mean(top1, top5)
	RareAct (Miech et al., 2020) [link]	2020	122	7,607	-	mWAP, mSAP
Object Detection	COCO 2014 Detection (Lin et al., 2014) [link]	2014	80	83,000	41,000	box mAP
	COCO 2017 Detection (Lin et al., 2014) [link]	2017	80	118,000	5,000	box mAP
	LVIS (Gupta et al., 2019) [link]	2019	1203	118,000	5,000	box mAP
	ODinW (Li et al., 2022) [link]	2022	314	132413	20070	box mAP
Semantic Segmentation	PASCAL VOC 2012 Segmentation (Everingham et al., 2010) [link]	2012	20	1464	1449	mIoU
	PASCAL Content (Mottaghi et al., 2014) [link]	2014	459	4998	5105	mIoU
	Cityscapes (Cordts et al., 2016) [link]	2016	19	2975	500	mIoU
	ADE20k (Zhou et al., 2017) [link]	2017	150	25574	2000	mIoU
Vase	VaseVQA	2025	7	9534	2339	Accuracy & BLEU

Table 7: Summary of the widely-used visual recognition datasets for MLLM evaluation.