ArtHopQA: Diagnosing Hidden-Chain Recovery in Artwork Question Answering

Xuefeng Wei[1], Xuan Zhou[1], Yuta Kato[2], Shintaro Ozaki[1], Kazuki Hayashi[1], Yusuke Sakai[1], Hidetaka Kamigaito[1], Katsuhiko Hayashi[2], Taro Watanabe[1]
[1]Nara Institute of Science and Technology (NAIST), Japan
[2]The University of Tokyo, Japan
{xuefeng.wei.yb1,xuan.zhou.xc7,ozaki.shintaro.ou6}@naist.ac.jp
{hayashi.kazuki.hl4,sakai.yusuke.sr9,kamigaito.h,taro}@is.naist.jp
{ukato6209,katsuhiko-hayashi}@g.ecc.u-tokyo.ac.jp

Abstract

Many artwork questions begin with vision but cannot be answered by visual recognition alone. A model may need to connect an artwork to its creator, depicted subject, collection, or location, and then follow multiple facts to reach the answer. We introduce ARTHOPQA, a multimodal multi-hop question answering benchmark for artwork understanding, containing 4,885 image-grounded questions paired with supporting fact chains. Experiments with open and API large vision-language models show a large gap: even the strongest open model reaches only 36.87% accuracy when given the image and title, while gold-path performance reaches 97.77% to 99.80% when the supporting fact chain is provided. Text-only language models exhibit a similar gap, showing that the difficulty is not confined to visual input. Further analysis reveals high-prior geographic guesses, inconsistent answers under image-title conflict, and limited recovery of intermediate entities even when final answers are correct.

1 Introduction

A vision-language model for artwork understanding should do more than name what appears in an image. When a user asks about an artwork, they may be asking for the historical, biographical, or institutional knowledge behind it. For example, given the artwork in Figure 1, the question *"Which country contains the birthplace of the creator of this artwork?"* cannot be answered by describing the scene. The model must connect the artwork to its creator, the creator to a birthplace, and the place to a country. The answer is therefore reached through a chain of facts: *artwork* → *creator* → *birthplace* → *country*. Such reasoning is central to artwork understanding, where images often serve as entry points to broader cultural and historical knowledge. Prior work on semantic artwork understanding and artwork explanation has shown that art interpretation requires knowledge beyond ordinary object recognition (Garcia and Vogiatzis, 2018; Hayashi et al., 2024).

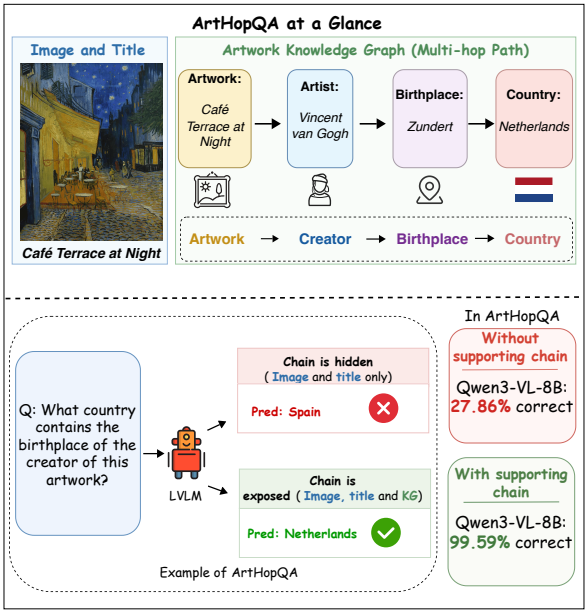


Figure 1: An ArtHopQA example illustrating multimodal multi-hop artwork QA. The question starts from an artwork image and title, but the answer is obtained only after following a multi-hop fact chain: artwork → creator → birthplace → country. The lower panel contrasts the same LVLM with and without this supporting chain, highlighting that the challenge is recovering the hidden multi-hop path from standard image/title input.

Large vision-language models (LVLMs) have made rapid progress in connecting images and language (Radford et al., 2021; Li et al., 2023; Liu et al., 2023), and knowledge-intensive visual question answering has shown that many visual questions require information beyond the image itself (Marino et al., 2019; Schwenk et al., 2022; Chen et al., 2023). However, artwork questions pose a distinctive challenge. Unlike many text-only multi-hop QA settings, the starting entity is not always provided in a clean textual form. The

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4057–4075
November 6-10, 2026 ©2026 Association for Computational Linguistics

model may need to infer the relevant artwork from an image or ambiguous title, then choose which relation chain among many artwork facts to follow. A knowledge graph stores such facts as relations between entities, and a multi-hop question requires following more than one relation (Yang et al., 2018; Trivedi et al., 2022; He et al., 2026). In artwork QA, the difficulty is not only combining multiple facts, but recovering which unseen chain is relevant from a visual or title-based entry point. We call this ability ***hidden-chain recovery***. It differs from standard VQA, where the answer may be directly visible, and from general multimodal QA, which may combine images with text, tables, or web evidence but does not focus on artwork-centered knowledge paths (Talmor et al., 2021; Chang et al., 2022).

We introduce ARTHOPQA[1], a multimodal multi-hop question answering benchmark for artwork understanding. Each example pairs an artwork image and its title with a question, an answer, and the supporting fact chain needed to derive that answer. The benchmark contains nearly 5,000 quality-controlled questions and a harder challenge split. The supporting chains are built from a Wikidata-derived artwork knowledge graph, where entities are represented with stable identifiers and questions use readable labels (Vrandečić and Krötzsch, 2014). We evaluate models under standard image/title inputs and evidence-exposed settings, where the supporting chain or local structured evidence is shown. These evidence-exposed settings are inspired by retrieval-augmented generation, but are used as diagnostics of whether models can use structured evidence once it is available, not as a claim that open-ended retrieval is solved (Lewis et al., 2020).

Our experiments reveal a large hidden-chain recovery gap. The strongest evaluated open LVLM, Qwen2.5-VL-72B, reaches 36.87% Norm-EM with Image+Title, compared with 99.80% when its gold supporting path is provided. Text-only language models exhibit a similar contrast between title-only and evidence-exposed inputs. Path-level diagnostics further distinguish final-answer correctness from recovery of the recorded intermediate entities. The benchmark and its executable supporting-path annotations are the primary contribution; the hidden-versus-exposed gap is an empirical finding enabled by this resource. Our empirical conclusions are specific to the artwork domain. Our contributions are as follows:

1. We introduce ArtHopQA, an artwork QA benchmark containing 4,885 image-grounded questions with executable two- or three-hop supporting fact chains.
2. We provide matched hidden-chain and evidence-exposed evaluation settings, together with path-level diagnostics, to examine answer prediction and supporting-chain recovery on the same examples.
3. We evaluate LVLMs and text-only language models, identifying limited intermediate-entity recovery and error patterns including high-prior geographic guesses, historical-to-modern collapses, and final-hop failures.

2 Related Work

Artwork understanding and art-language datasets. Prior work has explored how models connect artworks with language, semantics, and affective interpretation. Artpedia annotates paintings with both visual and contextual sentences, highlighting that artwork descriptions often involve information beyond visible content (Stefanini et al., 2019). ArtEmis studies emotional responses to artworks and provides natural-language explanations of why artworks evoke particular emotions (Achlioptas et al., 2021). These datasets show that artwork understanding is not limited to object recognition. However, they mainly support description, retrieval, affective interpretation, or explanation, rather than answerable questions whose answers can be traced through explicit multi-hop fact chains.

Knowledge-aware visual question answering. Several VQA benchmarks study questions that require external knowledge. FVQA pairs visual questions with supporting facts from external knowledge bases (Wang et al., 2018), while KVQA focuses on knowledge-aware questions involving named entities and multi-relation reasoning over knowledge graphs (Shah et al., 2019). These studies motivate the need to evaluate visual models with structured knowledge. However, most knowledge-aware VQA benchmarks are designed for general images or entity-centric visual questions. They do not specifically target artworks, where the image or title often serves as an entry point to art-historical, biographical, institutional, and geographic knowledge.

[1] https://github.com/Big-Sid/ArtHopQA

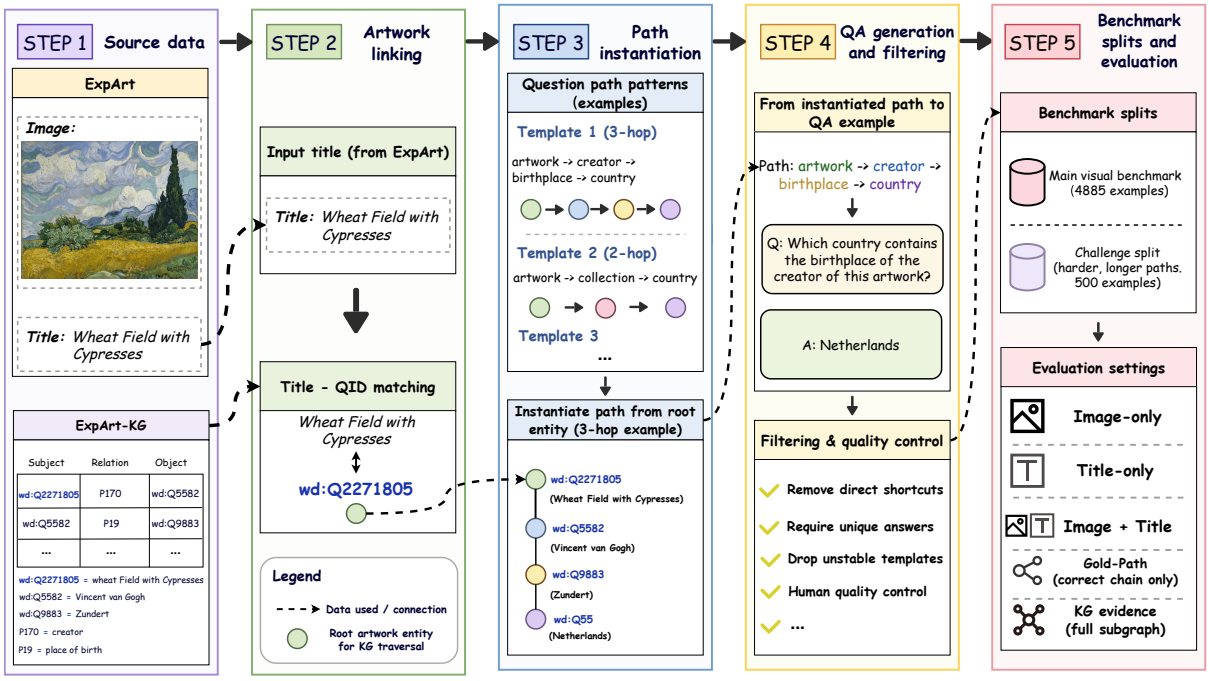


Figure 2: Overview of the ArtHopQA construction and diagnostic evaluation pipeline. We start from artwork images, titles from ExpArt, and ExpArt-KG. Artwork titles are linked to KG entities, allowed path patterns are instantiated from the linked artwork entity, and each resulting path is verbalized into a QA example with a stored supporting fact chain. After filtering and quality control, the benchmark supports standard image/title settings and evidence-exposed diagnostic settings.

Multi-hop and knowledge-graph question answering. Multi-hop QA and complex KBQA benchmarks have studied how models combine multiple pieces of evidence or execute compositional reasoning. 2WikiMultiHopQA provides reasoning paths for multi-hop questions constructed from structured and unstructured sources (Ho et al., 2020). ComplexWebQuestions studies decomposition of complex questions into simpler web-based questions (Talmor and Berant, 2018). KQA Pro provides explicit compositional programs and SPARQL queries for complex question answering over a knowledge base (Cao et al., 2022). These benchmarks make reasoning paths more explicit, but they are primarily text- or knowledge-base-centered. They do not study the setting where the reasoning path must be recovered from an artwork image or title before the model can answer.

Benchmark positioning. ArtHopQA combines an artwork image/title entry point, executable two- or three-hop supporting paths, and matched hidden-chain and evidence-exposed controls. It complements the heterogeneous evidence settings of WebQA and MultiModalQA (Chang et al., 2022; Talmor et al., 2021) with relation-level traces in a frozen artwork graph, and extends the path- or program-supervised evaluation perspective of 2WikiMultiHopQA and KQA Pro (Ho et al., 2020; Cao et al., 2022) to artwork-conditioned QA. The distinction is this combination of resource and diagnostic design, rather than the general observation that external evidence improves answering.

3 Task Definition

The gaps above motivate a benchmark that starts from artwork images and titles, but evaluates whether models can recover explicit multi-hop knowledge paths. We now define the ArtHopQA task.

Let $\mathcal{D}$ denote the benchmark. Each example $x \in \mathcal{D}$ is represented as

$$x = (v, t, q, a, p),$$

where v is the artwork image, t is the artwork title, q is the question, a is the gold answer label, and p is the supporting fact chain. The model predicts a textual answer $\hat{a}$, which is evaluated against a after normalization. We evaluate this task under six settings: three standard image/title settings that hide the supporting chain, and three evidence-

exposed settings that append flat KG triples, candidate paths, or the gold supporting path, as detailed in Section 5.2.

The chain p consists of k knowledge graph triples:

$$p = \big((e_0, r_1, e_1), (e_1, r_2, e_2), \ldots, (e_{k-1}, r_k, e_k)\big),$$

where $k \in \{2, 3\}$, e_0 is the artwork entity, and e_k is the answer entity associated with a. We use two- and three-hop paths to evaluate multi-hop knowledge recovery rather than direct attribute lookup: one-hop questions often reduce to recognition or memorized lookup, while longer paths are sparser and more ambiguous. Questions use readable labels, while supporting paths are stored with stable entity identifiers to preserve a traceable source of truth.

For example, a question about the country containing the birthplace of an artwork's creator may require the chain:

artwork → *creator* → *birthplace* → *country*.

In standard image/title settings, the model does not see this chain and must infer which path is relevant from the artwork image, title, or both.

4 Benchmark Construction

4.1 Construction Pipeline

Figure 2 summarizes the construction pipeline. **Step 1** collects two sources: ExpArt (Hayashi et al., 2024), an artwork explanation dataset that provides artwork images, titles, and metadata, and a frozen Wikidata-derived artwork knowledge graph introduced by Kato et al. (2026). We use ExpArt images and titles as the visual and textual entry points of our questions, and use the graph as the structured source for sampling supporting fact chains. The graph stores entity-relation triples, such as *(artwork, creator, artist)* and *(artist, place of birth, city)*. Statistics of the frozen source snapshot are reported in Appendix A.4. **Step 2** links each artwork title to an artwork entity, which becomes the root node for graph traversal. **Step 3** instantiates two- and three-hop paths from this root using relation templates, which define allowed path patterns over the graph. For example, the template *artwork* → *creator* → *birthplace* → *country* can yield the concrete path *Wheat Field with Cypresses* → *Vincent van Gogh* → *Zundert* → *Netherlands*. **Step 4** verbalizes each instantiated path with a deterministic question template, where the final entity becomes the answer. This design preserves alignment among the question, answer, and executable path, prioritizing traceability over linguistic diversity. Representative patterns and templates are listed in Appendix A.9. **Step 5** applies filtering and quality control to form the benchmark splits. Answer uniqueness is defined within the frozen graph, not across all external art-historical sources.

Quality audit. A two-annotator audit of 300 examples, followed by LLM-assisted and manually checked adjudication of disagreements, yields 98.00% answer validity (95% Wilson CI: 95.71–99.08) and 97.67% supporting-path validity (95.26–98.87). Appendices A.2 and A.3 detail the filtering and audit protocols.

4.2 Benchmark Splits

The main visual benchmark contains 4,885 image-grounded examples. Each example includes an artwork image, a title, a question, a gold answer, and a supporting fact chain. We use this split for all main experiments so that results across input and evidence settings are directly comparable.

We additionally provide a 500-example challenge split focused on collection-headquarters reasoning. These questions connect an artwork to its holding collection or institution and then to the institution's headquarters location, sometimes followed by the country of that location. The split is harder because headquarters information is rarely visible from the artwork input and has heterogeneous answer types, including cities, buildings, administrative locations, and institution-level locations. The challenge split is therefore not a noisier version of the main benchmark, but a valid harder split where the supporting chain is more difficult to recover from standard image/title inputs. Additional dataset statistics are reported in Appendix A.1.

5 Experimental Setup

We evaluate ArtHopQA as a diagnostic benchmark rather than a single leaderboard. The goal is to separate three abilities: using the artwork input, recovering the relevant multi-hop knowledge path, and producing the answer once the needed evidence is available. All main experiments use the 4,885-example visual benchmark, so results across input and evidence settings are directly comparable.

Setting	Input to model	Diagnostic role
Image-only	image, question	Visual entry
Title-only	title, question	Title-based entry
Image+title	image, title, question	Best standard input
Flat KG	input, question, unordered triples	Use exposed graph facts
Candidate path	input, question, candidate paths	Select from structured paths
Gold path	input, question, supporting chain	Oracle evidence use

Table 1: Evaluation settings. The first three settings hide the supporting chain, while the last three expose structured evidence for diagnostic purposes.

5.1 Models

We evaluate a representative set of open LVLMs, an API model, and text-only language models under the same concise-answer protocol. For open LVLMs, we evaluate Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), InternVL2.5-8B (Chen et al., 2025), and LLaVA-OneVision-7B (Li et al., 2024). For model-scale diagnostics, we additionally evaluate Qwen2.5-VL-3B, Qwen2.5-VL-32B, and Qwen2.5-VL-72B from the same Qwen2.5-VL family. We also evaluate GPT-5.4-mini as an API model reference system[2].

We further evaluate text-only language models, including Qwen2.5-7B-Instruct (Yang et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and DeepSeek-V3.2 (DeepSeek-AI et al., 2025). The text-only models receive the artwork title or structured evidence, but no image. They help test whether the gap is specific to visual perception or also appears when the task is reduced to text and evidence use.

5.2 Evaluation Settings

Table 1 summarizes the six evaluation settings. All prompts include the question and ask for a concise answer entity. Standard image/title settings hide the supporting chain, while evidence-exposed settings append structured evidence. Representative prompt formats and question-generation templates are provided in Appendices A.10 and A.9.

To make the evidence settings concrete, consider the question: *Which country contains the birthplace of the creator of this artwork?* For a Van Gogh artwork, the useful supporting chain may be *artwork* → *Vincent van Gogh* → *Zundert* → *Netherlands*. In the **flat KG** setting, the evidence block contains unordered local KG triples with concrete entity names and relation labels, such as *(artwork, creator, Vincent van Gogh)*, *(Vincent van Gogh, place of birth, Zundert)*, and *(Zundert, country, Netherlands)*, together with other nearby triples. Flat-KG prompts contain up to 100 triples, with a mean of 17.29 and a median of 15 triples per prompt.

In the **candidate-path** setting, the evidence block contains concrete multi-hop paths from the linked artwork entity, rather than abstract templates such as *artwork* → *creator* → *birthplace* → *country*. Candidate-path prompts contain up to 12 paths, with a mean of 6.75 and a median of 7 paths per prompt.

In the **gold-path** setting, the evidence block contains the single curated supporting path, so it is an oracle evidence setting. The non-oracle flat-KG and candidate-path settings do not use gold answer or gold-path metadata to select evidence, although the answer-bearing path may still appear because of high graph coverage; Section 7.1 tests this by removing the gold path. The main Flat-KG setting uses a whitelist of template relations and includes the stored gold path in 99.96% of prompts; the final Candidate-Path setting has 100.00% inclusion. These settings differ from the all-relation Flat-100 coverage ablation in Section 7.1, whose inclusion rate is 48.60%. Inclusion concerns the evidence displayed to the model: every benchmark example retains its gold answer and supporting path.

5.3 Metrics

We report normalized answer-label exact match, denoted as Norm-EM, as the primary metric. The normalization handles casing, punctuation, whitespace, Unicode quotes, and minor formatting. It does not collapse historically distinct entities. For example, Dutch Republic is not treated as Netherlands, and Republic of Venice is not treated as Italy. This distinction is important because historical polities are often intended answers in artwork questions.

All main benchmark comparisons use Norm-EM. Appendix A.12 additionally reports two diagnostic variants: *surface-alias matching*, which accepts documented surface aliases of the gold answer, and *occupation-relaxed matching*, which is applied only to `depicts_occupation_2` questions and accepts occupations of the depicted entity that

[2]GPT-5.4-mini: https://developers.openai.com/api/docs/models/gpt-5.4-mini. Accessed: 2026-05-25.

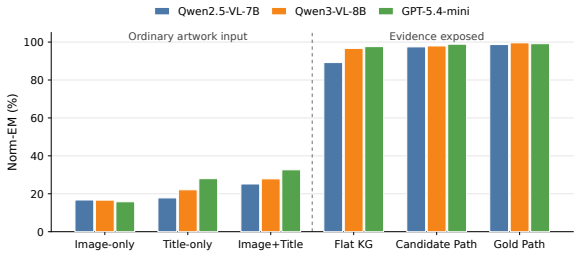


Figure 3: Performance ladder on the main visual benchmark. Standard image/title inputs remain low, while evidence-exposed settings sharply improve performance. Flat KG and candidate-path settings expose structured evidence and are interpreted as evidence-exposure diagnostics.

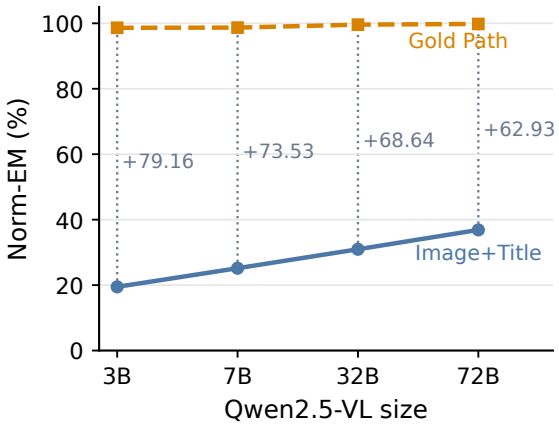


Figure 4: Scaling trend for Qwen2.5-VL. Image+title accuracy improves with model size, but remains far below gold-path performance, indicating that scaling alone does not close the hidden-chain recovery gap.

appear in the frozen local KG. For selected main comparisons, including image+title versus single-modality inputs, gold-path versus image+title, and selected evidence-exposed comparisons, we report bootstrap confidence intervals and paired comparisons in Appendix A.6.

Semantic sensitivity. For Qwen2.5-VL-72B, we audit 250 randomly sampled Norm-EM-negative Image+Title outputs using a two-stage LLM-judge protocol. The audit identifies 13 additional semantically acceptable answers, giving an estimated adjusted accuracy of 40.15% (95% CI: 38.80–42.36), compared with 36.87% under Norm-EM. This remains far below its 99.80% Gold-Path Norm-EM. This supplementary estimate does not replace the official metric; Appendix A.12 details the sampling, repeated same-model adjudications, and interval calculation.

6 Results

We organize the results as a diagnostic ladder. We first compare standard image/title inputs with evidence-exposed settings on the main visual benchmark. We then examine whether model scaling and the challenge split change the same pattern.

6.1 The Hidden-Chain Gap Across Models

Figure 3 shows the main pattern for representative models. Standard image/title inputs remain low, while evidence-exposed settings sharply improve performance. The figure illustrates the central gap: models perform poorly when they must recover the supporting chain from the image and title, but perform much better when structured evidence is exposed.

Table 2 gives the full diagnostic matrix. Across open LVLMs, image+title performance remains far below gold-path performance. Scaling improves standard inputs, but even Qwen2.5-VL-72B reaches only 36.87% with image+title, compared with 99.80% when the gold path is provided. GPT-5.4-mini narrows the gap relative to smaller open LVLMs, but its image+title score also remains far below its gold-path score.

The evidence-exposed columns identify the bottleneck more directly. Flat KG and candidate-path evidence sharply improve most models, and text-only language models show the same pattern: title-only inputs are weak, while structured evidence is much stronger. LLaVA-OneVision improves less under noisy KG evidence than other LVLMs, but its gold-path score remains high, suggesting weaker evidence selection rather than unanswerable questions.

6.2 Scale Helps but Does Not Close the Gap

Figure 4 isolates the Qwen2.5-VL scale trend. Image+title performance improves from 19.47% at 3B to 36.87% at 72B. However, gold-path performance stays near saturation across scales. Therefore, scaling narrows the gap but does not solve the hidden-chain recovery problem.

6.3 The Challenge Split Sharpens the Gap

The 500-example challenge split isolates collection-headquarters reasoning. These questions connect an artwork to its holding collection or institution and then to the institution's headquarters location, sometimes followed by the country of that location. This relation family is harder because the head-

Model	Standard input			Evidence exposed			Gap
	Image	Title	Image+Title	Flat KG	Cand. Path	Gold Path	
Open LVLMs							
Qwen2.5-VL-3B	14.39	15.84	19.47	85.71	91.18	98.63	**79.16**
Qwen2.5-VL-7B	16.72	17.81	25.16	89.19	97.46	98.69	**73.53**
Qwen2.5-VL-32B	17.48	25.30	30.95	96.40	98.46	99.59	**68.64**
Qwen2.5-VL-72B	22.64	31.03	36.87	98.32	99.24	99.80	**62.93**
Qwen3-VL-8B	16.64	22.09	27.86	96.66	97.95	99.59	**71.73**
InternVL2.5-8B	16.99	22.21	27.82	93.76	91.93	99.45	**71.63**
LLaVA-OneVision-7B	17.26	18.89	24.24	40.51	53.41	97.77	**73.53**
API LVLM reference							
GPT-5.4-mini	15.78	27.96	32.63	97.60	98.87	99.14	**66.51**
Text-only language models							
Qwen2.5-7B-Instruct	–	20.14	–	93.29	98.92	98.49	**78.35**
Llama-3.1-8B-Instruct	–	16.95	–	82.68	84.01	96.93	**79.98**
DeepSeek-V3.2	–	27.72	–	95.68	98.01	99.67	**71.95**

Table 2: Unified diagnostic results on ArtHopQA. Values are Norm-EM. For LVLMs, Gap is Gold Path minus Image+Title; for text-only language models, Gap is Gold Path minus Title. Bold Gap values highlight the hidden-chain recovery gap. Flat KG and candidate-path settings expose structured evidence from the frozen local graph and should be interpreted as evidence-exposure diagnostics, not open-ended retrieval results. Dashes indicate settings that are not applicable to text-only models.

Model	I+T	Gold	Gap	95% CI
Qwen2.5-VL-7B	8.20	98.40	90.20	[87.40, 92.80]
Qwen3-VL-8B	11.60	100.00	88.40	[85.60, 91.20]
InternVL2.5-8B	14.00	100.00	86.00	[83.00, 89.20]
LLaVA-OV-7B	11.20	95.60	84.40	[81.20, 87.40]
GPT-5.4-mini	19.00	100.00	81.00	[77.40, 84.80]

Table 3: Challenge-split Norm-EM (%; $n = 500$). I+T denotes Image+Title; LLaVA-OV denotes LLaVA-OneVision. Gap is Gold-Path minus I+T in percentage points, with paired 95% bootstrap CIs (1,000 resamples; Appendix A.6).

Evidence	Gold path incl. (%)	Qwen2.5	Qwen3
Cand. top-10	100.00	97.15	99.69
Cand. no-gold@10	0.00	54.88	61.74
All-rel. Flat-100	48.60	58.14	69.85
All-rel. Flat no-gold	0.00	19.61	27.96

Table 4: Evidence ablations on 4,885 examples (Norm-EM, %). Gold-path inclusion is the fraction of prompts containing the stored supporting path; "All-rel." denotes the all-relation evidence builder.

quarters entity is rarely indicated by the artwork image or title, and its answer type is heterogeneous, ranging from cities and buildings to administrative or institution-level locations. Table 3 shows that standard image+title inputs are substantially lower on this split, while gold-path performance remains near saturation. This suggests that the challenge split is difficult because the relevant chain is harder to recover and type correctly, not because the questions are unanswerable.

Overall, the results show a consistent pattern. Images, titles, model scale, and stronger API systems provide useful signal, but they do not remove the gap between standard image/title input and explicit supporting evidence. The next section analyzes why this gap persists and whether it is robust to alternative explanations.

7 Analysis

The results show a consistent gap between standard image/title inputs and evidence-exposed settings. We now analyze whether this gap is robust to alternative explanations and where models fail when the supporting chain is hidden.

7.1 Correct Evidence Is Crucial

The evidence-exposed settings test whether models can use structured graph evidence. Table 4 examines their dependence on the answer-bearing path. Candidate top-10 reaches 97.15% and 99.69% for Qwen2.5-VL-7B and Qwen3-VL-8B, respectively; removing the gold path reduces these scores to 54.88% and 61.74%. The all-relation Flat-100 ablation also loses substantial accuracy after removing gold-path triples. This ablation uses a different evidence builder from the whitelisted Flat-KG setting in the main results.

Model	Original	Para.	Delta
Qwen2.5-VL-7B	29.90	30.20	+0.30
Qwen3-VL-8B	31.70	31.80	+0.10

Table 5: Paraphrase robustness on 1,000 examples (Norm-EM, %).

Setting	Qwen2.5	Qwen3
Image–title conflict	83.66	81.00
Image swap	78.67	73.16

Table 6: Neither-consistency under counterfactual inputs (%): predictions matching neither the image-side nor title-side reference answer.

Model	Final	Ent. R	All+Ans	Strict
Qwen2.5	24.20	7.41	1.00	0.00
Qwen3	34.40	26.43	10.50	0.00

Table 7: Path diagnostics on 1,000 stored traces per model (%). Ent. R is alias-aware gold-step entity recall; All+Ans requires all gold step entities and a Norm-EM-correct answer; Strict additionally requires the correct final answer together with positionally matched step entities and relation labels.

The residual no-gold performance can reflect other graph context or model knowledge, but near-saturated performance depends strongly on access to answer-bearing evidence. Appendix A.5 reports the full coverage ablations.

7.2 Template Dependence and Answer Shortcuts

Because ArtHopQA uses template-based questions, we test both sensitivity to surface wording and template-level answer shortcuts. Table 5 reports rule-based paraphrases on 1,000 examples. Aggregate accuracy changes by +0.30 and +0.10 percentage points for Qwen2.5-VL-7B and Qwen3-VL-8B, respectively, although individual predictions can change. Appendix A.5 reports both subset sizes and the lost/gained rates.

Template-level answer shortcuts. On all 4,885 examples, a per-template majority-answer baseline reaches 4.14% Norm-EM when the answer is selected using all evaluation labels, and 1.17% under stratified five-fold evaluation. Removing both the image and title gives question-only scores of 1.86% for Qwen2.5-7B-Instruct and 0.00% for Llama-3.1-8B-Instruct. This leaves 14 unique prompts, one per template; deterministic evaluation decodes each once and applies its prediction to the corresponding examples. These controls test template-level fixed-answer priors, not generalization to human-authored questions or unseen templates.

7.3 Answer Consistency under Image-Title Conflict

We also test how models behave when the image and title point to different artworks. We compare each prediction with the image-side and title-side reference answers. Most predictions match neither reference.

Table 6 summarizes the key neither-consistency rates. In image-title conflict examples, Qwen2.5-VL-7B produces neither-consistent answers in 83.66% of cases, and Qwen3-VL-8B does so in 81.00% of cases. Image-swap tests show a similar pattern. Appendix A.5 reports the full counterfactual results, including image-consistent and title-consistent rates.

Under conflicting cues, predictions are not consistently aligned with either reference answer. Because answering also requires background knowledge, neither-consistency alone does not isolate a perceptual grounding failure.

7.4 Final Answers Do Not Imply Faithful Path Recovery

Final-answer accuracy does not establish whether a model has recovered the recorded supporting path. We ask models to output intermediate entities, relation labels, and the final answer, and additionally rescore the same stored traces using alias-aware entity diagnostics.

Table 7 compares strict and relaxed results on 1,000 traces per model. For Qwen2.5-VL-7B, all traces are parseable, yet gold-step entity recall is 7.41%, and all gold entities plus the correct final answer are recovered on only 1.00% of examples. Qwen3-VL-8B has higher entity recall (26.43%) and joint recovery (10.50%), but both remain below its 34.40% final-answer Norm-EM. Limited recovery therefore persists without exact relation-label matching. Appendix A.5 reports the original strict diagnostics, relaxed precision/recall/F1 for both models, Qwen2.5-VL-7B hop-wise recovery, and a 100-trace scorer validation.

Table 8 illustrates a correct answer with an incorrect intermediate collection, an incorrect creator followed by an incorrect answer, and a successful reported entity sequence. These examples concern

The Archangel Michael	
Gold	Benaki Museum → Greece
Reported	National Gallery of Greece → Greece
Return to the Sheepfold	
Gold	Carlo Pittara → Kingdom of Italy
Reported	Jean-François Millet → France
The Knifegrinder	
Gold	Francisco Goya → Spain → Madrid
Reported	Francisco Goya → Spain → Madrid

Table 8: Representative reported entity sequences. The shared artwork root is omitted. A correct final answer can coexist with an incorrect intermediate entity, while the tracing protocol can also record a correct entity sequence.

Error category	**Share**
High-prior geographic guess	47.35
True unrelated	37.41
Wrong relation selected	9.74
Historical-to-modern collapse	1.56
Final-hop missing	1.51

Table 9: Top semantic error categories over wrong prediction rows. Values are percentages.

the correctness of reported traces, not a direct observation of internal model computation.

7.5 Error Taxonomy

We further categorize wrong predictions to understand common failure modes. Table 9 reports the top error categories. The largest category is high-prior geographic guessing, accounting for 47.35% of wrong prediction rows. In these cases, models output frequent art-historical countries or cities despite the supporting path pointing elsewhere. The next largest category is unrelated answers, followed by wrong relation selection. Appendix A.5 reports the complete taxonomy, including historical-to-modern collapses, final-hop failures, surface aliases, and answer-type mismatches.

These errors suggest that models often produce plausible art-related knowledge, but fail to follow the exact multi-hop path required by the question. This aligns with the quantitative results: models can often use the supporting chain once it is shown, but they struggle to recover it from standard image/title inputs.

8 Conclusion

We introduced ArtHopQA, an artwork QA benchmark with executable supporting paths and matched hidden-chain and evidence-exposed evaluations. Across the evaluated models, standard artwork inputs remain substantially below evidence-exposed settings. Template-level controls and a semantic-scoring sensitivity analysis support this contrast, while relaxed path diagnostics and stored examples distinguish correct final answers from correct intermediate-entity recovery. The benchmark provides a traceable resource for evaluating artwork-conditioned knowledge recovery and evidence use.

Limitations

Template-based questions. ArtHopQA uses relation-specific templates to preserve alignment among questions, answers, and supporting paths. The paraphrase and question-only controls test surface wording and template-level answer priors, but do not evaluate generalization to unseen relation templates or naturally authored user questions. Human-authored, open-ended, and conversational artwork questions remain complementary extensions.

Frozen knowledge graph. The supporting paths are sampled from a frozen Wikidata-derived artwork knowledge graph. This design gives each answer a traceable source of truth, but the graph may have incomplete coverage, outdated facts, or uneven representation across artists, regions, historical periods, and institutions. Some art-historical entities, such as historical polities, disputed attributions, or changing collection locations, can also be sensitive to source conventions. Our empirical conclusions are specific to this artwork-domain benchmark and do not establish cross-domain or cross-KG transfer.

Evidence-exposed settings. The evidence-exposed settings should not be interpreted as end-to-end retrieval systems. The gold-path setting is an oracle diagnostic, and the KG-evidence settings expose local structured evidence from the frozen graph. These settings are useful for identifying whether models can use relevant evidence once it is available, but they do not solve the open problem of retrieving the correct path from arbitrary artwork inputs.

Visual grounding. The benchmark is visually grounded through artwork images and titles, but not every question requires fine-grained visual perception. The task is best understood as artwork-conditioned multi-hop knowledge recovery: the

image and title provide the entry point, while the answer often depends on background knowledge stored in the graph. Future work could create subsets that require more direct visual recognition, such as identifying depicted entities or distinguishing visually similar artworks.

Language and evaluation. The benchmark focuses on English templates and exact-match-style answer evaluation. Normalization, alias diagnostics, and the LLM-judge sensitivity analysis address some surface variation, but the latter audits sampled negatives for one model and uses repeated sessions of the same judge model. It does not replace independent expert adjudication or establish semantic accuracy for every evaluated system. Multilingual questions and richer answer and explanation evaluation remain future extensions.

Ethical Considerations

Data release. The ArtHopQA project repository is publicly available at https://github.com/Big-Sid/ArtHopQA. The release includes benchmark examples, questions, answers, supporting paths, metadata, evaluation scripts, and prompts for reproducing the reported results, together with documentation of the construction, filtering, and quality-control procedures.

Use of AI assistants. AI assistants were used primarily for language editing and manuscript proofreading during manuscript preparation, e.g., the camera-ready phase. All scientific content and final revisions were reviewed and approved by the authors.

Licenses and source resources. ArtHopQA is built from publicly available artwork resources and a Wikidata-derived knowledge graph. ExpArt, which provides artwork images and titles used as visual entry points, is released on Hugging Face under a CC BY-SA 4.0 license. Wikidata structured data from the main, Property, Lexeme, and EntitySchema namespaces is available under CC0. However, individual media files linked from Wikimedia or Wikipedia may have their own licenses and usage conditions. For this reason, our release will provide metadata, image URLs, and retrieval instructions rather than claiming ownership over source images.

Cultural and historical bias. Artwork datasets and knowledge graphs may reflect historical, geographic, linguistic, and institutional biases. Coverage may be stronger for well-documented European artworks and major museum collections than for less digitized artistic traditions. Model performance on ArtHopQA should therefore not be interpreted as balanced coverage of all cultures, periods, or artistic communities.

Art-historical sensitivity. Knowledge-graph labels may simplify complex art-historical issues, including disputed authorship, historical geography, colonial collections, religious or mythological entities, and changing institutional locations. We keep supporting paths traceable to stable entity identifiers to improve transparency, but the benchmark should not be used as an authoritative source for art-historical judgments without expert review.

Intended use. The benchmark is intended for research on multimodal reasoning, knowledge-grounded evaluation, and artwork understanding. It is not designed for high-stakes decisions, cultural heritage attribution, provenance claims, or expert-level museum cataloging. Researchers using ArtHopQA should be careful when deploying artwork QA systems in public-facing settings, since incorrect answers about cultural heritage may mislead users or reinforce existing biases.

Acknowledgments

This work was partly supported by JSPS Kakenhi Grant Number JP26H02537 and was also supported by the China Scholarship Council (CSC) under grant numbers 202308070070 (Xuan Zhou) and 202308070072 (Xuefeng Wei)

References

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. 2021. Artemis: Affective language for visual art. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11564–11574.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei

Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. 2022. KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6101–6119, Dublin, Ireland. Association for Computational Linguistics.

Yingshan Chang, Guihong Cao, Mridu Narang, Jianfeng Gao, Hisami Suzuki, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16474–16483.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Noa Garcia and George Vogiatzis. 2018. How to read paintings: Semantic art understanding with multi-modal retrieval. In *Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part II*, page 676–691, Berlin, Heidelberg. Springer-Verlag.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards artwork explanation in large-scale vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 705–729, Bangkok, Thailand. Association for Computational Linguistics.

Jie He, Nan Hu, Wanqiu Long, Jiaoyan Chen, and Jeff Z. Pan. 2026. MINTQA: A multi-hop question answering benchmark for evaluating LLMs on new and long-tail knowledge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 445–479, San Diego, California, United States. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuta Kato, Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2026. ExpArt-KG: Artwork Image Description Generation through Iterative Exploration of Knowledge Graphs. ResearchGate. DOI: 10.13140/RG.2.2.29734.08003.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

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. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, page 146–162, Berlin, Heidelberg. Springer-Verlag.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: knowledge-aware visual question answering. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2019. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In *Image Analysis and Processing – ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II*, page 729–740, Berlin, Heidelberg. Springer-Verlag.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *Preprint*, arXiv:2104.06039.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. ♫ MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2413–2427.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

A.1 Dataset Statistics

Table 10 summarizes the benchmark statistics and quality-control outcomes.

Statistic	Count
Visual evaluation split	4,885
Challenge split	500
Quality audit sample	300
Usable / Flagged / Removed	284 / 13 / 3
Final-action agreement	95.67%
Issue-presence agreement	96.67%

Table 10: ArtHopQA statistics and quality-control summary.

A.2 Construction Filters

Table 11 summarizes the main automatic filters used during benchmark construction. These filters reduce shortcut examples, answer ambiguity, generic answers, and unstable relation patterns. The challenge split is separated from the main benchmark because collection-headquarters relations have more heterogeneous answer types.

Filter	Purpose
Direct shortcut	Remove one-hop leakage
Unique answer	Reduce ambiguity
Repeated entities	Remove cyclic paths
Depicted entity	Disambiguate depicted people
Answer blacklist	Remove generic answers
Template cleanup	Remove unstable patterns
Challenge split	Separate heterogeneous relations

Table 11: Construction filters.

A.3 Quality Audit Schema and Agreement

We conduct a two-annotator quality audit on 300 examples sampled from the main benchmark. Each

annotator checks five aspects of an example: question naturalness, answer correctness, whether the question requires multi-hop reasoning, supporting-path correctness, and image-title consistency.

Action labels. Annotators assign one of three final actions:

- **KEEP**: the example is directly usable.
- **REVISE**: the example is mostly valid but should be flagged for wording, metadata, or borderline art-historical issues.
- **DROP**: the example has a hard failure, such as an incorrect path, wrong depicted entity, or title-path mismatch.

For readability in the main paper, we refer to these as usable, flagged, and removed examples. The final counts are 284 KEEP, 13 REVISE, and 3 DROP.

Agreement. Table 12 reports inter-annotator agreement. For the final KEEP/REVISE/DROP decision, annotators achieve 95.67% raw agreement with Cohen's $\kappa = 0.561$. For binary issue presence, they achieve 96.67% raw agreement with $\kappa = 0.670$. Some boolean fields have low κ despite high raw agreement because hard failures are rare, which creates strong label imbalance. A dash indicates that Cohen's κ is undefined because both annotators assigned the same label to all examples.

QC item	Agreement	Cohen's κ
Question naturalness	96.67%	0.529
Answer correctness	97.67%	0.212
Requires multi-hop	100.00%	–
Path correctness	97.33%	0.189
Image-title match	100.00%	–
Final action	95.67%	0.561
Issue presence	96.67%	0.670

Table 12: Two-annotator quality audit on 300 examples. A dash indicates that Cohen's κ is undefined because both annotators assigned the same label to all examples. Low κ for some boolean checks reflects rare negative labels despite high raw agreement.

Disagreements. Most disagreements are between KEEP and REVISE rather than KEEP and DROP. These cases usually involve borderline metadata or art-historical issues, such as disputed attribution, historical geography, low-salience depicted entities, or role-like occupation labels. This suggests that disagreements mainly concern whether an example should be flagged, rather than whether it is a hard failure.

Adjudication. We collect the union of 14 examples with disagreements involving answer correctness, supporting-path correctness, or final action. These examples are subjected to blind LLM-assisted adjudication using the artwork image, question, gold answer, and supporting path, without revealing the earlier annotator labels. The resulting decisions are subsequently checked manually against the same evidence, with no remaining discrepancies.

After adjudication, answer validity is 98.00% (95% Wilson CI: 95.71–99.08) and supporting-path validity is 97.67% (95.26–98.87). No answer/path cases remain unresolved in the 300-example audit.

A.4 Source Snapshot Statistics

Table 13 summarizes the frozen artwork source snapshot used for ArtHopQA path sampling. The snapshot combines ExpArt artwork images and titles with a Wikidata-derived artwork knowledge graph. We use this graph only as the structured source for sampling supporting fact chains; KG construction is not a contribution of this paper.

Statistic	Count
ExpArt rows	19,181
Unique artwork titles	7,269
Unique image URLs	7,260
High-confidence title–QID links	7,227
Linked titles with outgoing KG edges	7,034
Unique QID nodes	39,887
Unique QID triples	300,595
Unique relation PIDs	643

Table 13: Statistics of the frozen artwork source snapshot used for ArtHopQA path sampling.

A.5 Additional Diagnostic Results

This subsection provides the complete numerical results for the diagnostics discussed in Section 7. The main text reports compact supporting tables for each analysis claim, while the appendix provides the full evidence ablation grid, paraphrase robustness results, visual counterfactual results, path-level tracing results, and error taxonomy.

Evidence ablations. Table 14 tests whether high evidence-exposed scores depend on the correct supporting path being present. Candidate-path settings perform near the gold-path upper bound when the gold path is included. When the gold path is removed, performance drops sharply. This supports our interpretation that evidence-exposed

Mode	Gold-path incl. (%)	Paths	Qwen2.5	Qwen3
Cand top1	98.12	1.00	98.96	99.84
Cand top3	100.00	2.91	97.38	99.71
Cand top5	100.00	4.49	97.46	99.69
Cand top10	100.00	6.55	97.15	99.69
Cand no-gold	0.00	5.78	54.82	61.80
Cand no-gold@10	0.00	5.68	54.88	61.74
Distractor only	0.00	5.68	54.88	61.74
All-rel. Flat-20	17.32	0.00	39.49	49.93
All-rel. Flat-50	36.52	0.00	50.71	63.07
All-rel. Flat-100	48.60	0.00	58.14	69.85
All-rel. Flat no-gold	0.00	0.00	19.61	27.96

Table 14: Evidence ablation results on the 4,885-example visual benchmark. Values are Norm-EM except for gold-path inclusion and average number of paths. "Cand" denotes candidate-path evidence. "All-rel. Flat" denotes the all-relation coverage-ablation builder, which differs from the whitelisted Flat-KG builder used in the main diagnostic evaluation. Qwen2.5 and Qwen3 refer to Qwen2.5-VL-7B and Qwen3-VL-8B.

Model	N	Orig.	Para.	Delta
Qwen2.5-VL-7B	300	28.67	30.67	+2.00
Qwen3-VL-8B	300	32.67	31.00	-1.67
Qwen2.5-VL-7B	1000	29.90	30.20	+0.30
Qwen3-VL-8B	1000	31.70	31.80	+0.10

Table 15: Paraphrase robustness results. Values are Norm-EM on sampled examples. Orig. denotes the original template question, and Para. denotes the rule-based paraphrase.

gains come from exposing the answer-bearing path, rather than arbitrary KG context.

Paraphrase robustness. Table 15 reports rule-based paraphrase robustness. The 1,000-example subset shows changes below one point for both models. Aggregate accuracy is stable under the tested rule-based paraphrases, although individual predictions can change.

For the 1,000-example subset, paraphrasing changes individual predictions more often than the small aggregate deltas alone suggest. For Qwen2.5-VL-7B, 4.30% of examples change from correct to incorrect and 4.60% change from incorrect to correct. For Qwen3-VL-8B, the corresponding rates are 4.00% and 4.10%. Thus, aggregate accuracy is stable under the tested paraphrases even though some individual predictions change.

Template-level shortcut controls. Across all 4,885 examples, the per-template majority-answer baseline reaches 4.14% Norm-EM (95% CI: 3.61–4.73) when the majority answer is selected using all evaluation labels. Under stratified five-fold evaluation, the corresponding score is 1.17% (0.90–1.51).

For the question-only setting, Qwen2.5-7B-Instruct obtains 1.86% (1.52–2.28), compared with 20.14% (19.04–21.29) in the title-only setting. Llama-3.1-8B-Instruct obtains 0.00% (0.00–0.08) question-only, compared with 16.95% (15.92–18.03) title-only. Removing the artwork input leaves 14 unique prompts, one per template; under deterministic decoding, each prompt is decoded once and its prediction is applied to all examples sharing that template. Intervals are Wilson 95% confidence intervals.

Model	Set	Visual	Title	Neither
Qwen2.5	Conflict	10.60	5.73	83.66
Qwen3	Conflict	8.74	10.26	81.00
Qwen2.5	Image swap	5.77	15.56	78.67
Qwen3	Image swap	5.98	20.86	73.16

Table 16: Visual counterfactual results. Values are percentages. Visual denotes predictions consistent with the image-side answer, Title denotes predictions consistent with the title-side answer, and Neither denotes predictions matching neither side. Qwen2.5 and Qwen3 refer to Qwen2.5-VL-7B and Qwen3-VL-8B.

Visual counterfactuals. Table 16 reports image-title conflict and image-swap tests. When image and title cues conflict, most predictions match neither the image-side answer nor the title-side answer. These results show that predictions are not consistently aligned with either reference answer under conflicting cues. Because answering also requires background knowledge, the neither rate does not by itself isolate a perceptual grounding failure.

Path-level tracing. Table 17 reports a strict path-level diagnostic. Models are asked to output intermediate entities, relation labels, and the final answer. Although final-answer accuracy is higher than first-hop accuracy, exact path match is zero under this strict positional criterion. This suggests that final-answer correctness does not imply faithful recovery of the supporting path.

Relaxed entity diagnostics. We additionally rescore the same 1,000 stored traces using alias-aware entity matching without requiring an exact relation-label sequence. For Qwen2.5-VL-7B, entity precision/recall/F1 is 9.64/7.41/8.38%, and all gold step entities together with a Norm-EM-correct final answer are recovered on 1.00% of examples. For Qwen3-VL-8B, the correspond-

Model	N	Parse	Final	Hop1	Path
Qwen2.5-VL-7B	500	100.00	22.40	11.00	0.00
Qwen3-VL-8B	500	100.00	33.80	18.80	0.00
Qwen2.5-VL-7B	1000	100.00	24.20	9.90	0.00
Qwen3-VL-8B	1000	99.90	34.40	15.60	0.00

Table 17: Path-level tracing results. Values are percentages. Final denotes final-answer Norm-EM under the path-tracing prompt, Hop1 denotes first-hop entity accuracy, and Path denotes exact path match. Path exact requires the final answer, every step entity, and every relation label to match positionally.

ing precision/recall/F1 is 22.09/26.43/24.06%, and joint recovery reaches 10.50%. For Qwen2.5-VL-7B, relaxed positional Hop-1/2/3 entity recovery is 13.40/1.00/0.00%, with Hop-3 evaluated only on applicable three-hop examples. These relaxed diagnostics retain every example in the denominator and do not require an exact relation-label sequence.

Relaxed-scorer validation. We validate the relaxed scorer on 100 stored traces, with 50 examples from each model, stratified by automatic complete-chain and final-answer status. Two separately shuffled, blind GPT-5.6 Sol Pro sessions judge semantic entity equivalence without seeing the automatic labels or model identity. The two adjudication sessions agree on all 441 non-NA judgments (100.00%; Wilson 95% CI: 99.14–100.00; Cohen's $\kappa = 1.000$). Against this consensus, the automatic scorer obtains 95.24/94.12/94.67 precision/recall/F1 over 241 existing hop-entity judgments, 100.00/96.30/98.11 over 100 complete ordered entity-chain judgments, and 100.00/96.88/98.41 over 100 final-answer equivalence judgments. Because rare automatic complete-chain positives are intentionally oversampled, this stratified sample validates the scorer rather than estimating corpus-level path accuracy.

Error taxonomy. Table 18 reports the semantic error taxonomy over wrong prediction rows. The largest category is high-prior geographic guessing, where models predict common art-historical countries or cities despite the supporting path pointing elsewhere. Other notable errors include wrong relation selection, historical-to-modern collapses, and final-hop failures.

Error category	Share
High-prior geographic guess	47.35
True unrelated	37.41
Wrong relation selected	9.74
Historical-to-modern collapse	1.56
Final-hop missing	1.51
Valid surface alias	1.35
Answer type mismatch	1.08

Table 18: Semantic error taxonomy over wrong prediction rows. Values are percentages.

Setting	Norm-EM	CI low	CI high
Qwen2.5-VL-7B image+title	25.16	23.99	26.47
Qwen2.5-VL-7B gold-path	98.69	98.36	99.02
Qwen3-VL-8B image+title	27.86	26.63	29.21
Qwen3-VL-8B gold-path	99.59	99.41	99.75
Llama-3.1-8B title-only	16.95	15.95	18.04
Llama-3.1-8B candidate-path	84.01	82.93	84.99

Table 19: Selected bootstrap confidence intervals for main comparisons. Values are Norm-EM percentages.

A.6 Bootstrap Confidence Intervals and Paired Comparisons

We report bootstrap confidence intervals over benchmark examples for selected main comparisons. These intervals quantify example-level uncertainty for fixed pretrained models under deterministic decoding. Because no model training is involved, we report bootstrap intervals over examples rather than multi-seed training variance.

Challenge-split uncertainty. Table 20 reports the full challenge-split bootstrap intervals corresponding to Table 3.

A.7 Error Taxonomy Definitions

Table 22 defines the semantic error categories used in Section 7.5. The categories are assigned by matching wrong predictions against the gold answer, intermediate entities in the supporting path, relation-specific alternatives, historical-entity mappings, and high-prior geographic answer lists.

A.8 Model, API, and Decoding Details

Model identifiers. Table 23 lists the model identifiers and access modes used in our experiments. For open models, we use released checkpoints or their corresponding serving endpoints. For GPT-5.4-mini, we use the official OpenAI API model identifier `gpt-5.4-mini`. We report it as an API LVLM reference system because its training data, parameter count, and serving details are not fully transparent.

Model	I+T	Gold	Gap
Qwen2.5-VL-7B	8.20 [6.00, 10.60]	98.40 [97.20, 99.40]	90.20 [87.40, 92.80]
Qwen3-VL-8B	11.60 [8.80, 14.60]	100.00 [100.00, 100.00]	88.40 [85.60, 91.20]
InternVL2.5-8B	14.00 [11.20, 17.40]	100.00 [100.00, 100.00]	86.00 [83.00, 89.20]
LLaVA-OV-7B	11.20 [8.40, 14.20]	95.60 [93.80, 97.20]	84.40 [81.20, 87.40]
GPT-5.4-mini	19.00 [15.40, 22.60]	100.00 [100.00, 100.00]	81.00 [77.40, 84.80]

Table 20: Challenge-split Norm-EM and paired-gap estimates with 95% bootstrap CIs from 1,000 resamples of the same 500 examples.

Decoding and prompts. Unless otherwise noted, we use deterministic decoding with temperature $= 0$, top-p $= 1$, and one response per prompt. Prompts ask the model to return a concise answer entity. For standard image/title settings, the prompt includes the question and the relevant image and/or title. For evidence-exposed settings, the prompt additionally includes flat KG triples, candidate paths, or the gold supporting path. We do not fine-tune any model.

API usage and cost reporting. For GPT-5.4-mini, API calls are made through the official OpenAI API using `gpt-5.4-mini`. We report the exact model identifier and access date because API model implementations and pricing can change over time. Cost estimates, when reported, are computed from logged token usage and the official OpenAI pricing page at the time of evaluation.[3] We do not use API cost as an evaluation metric; it is reported only for reproducibility and responsible-experiment documentation.

Annotator details. The quality audit was conducted by two project-affiliated annotators from East-Asia. Both were familiar with the annotation schema and the benchmark construction process. The audit did not involve crowdworkers or collection of personal data. Annotators reviewed only benchmark examples, artwork metadata, questions, answers, and supporting paths. No private or sensitive personal information was collected. If required by the reviewing venue, this quality-control activity should be treated as internal dataset validation rather than a human-subjects study.

[3] https://openai.com/api/pricing; accessed May 2026.

Evaluation implementation. All reported main scores use the same normalized exact-match scorer. The scorer lowercases text, normalizes Unicode quotes and whitespace, removes punctuation and minor formatting artifacts, and compares against the canonical answer label. The main scorer does not collapse historically distinct entities. Alias-relaxed and occupation-relaxed diagnostics are computed only as supplementary analyses.

Compute budget. All experiments are inference-only; no model is trained or fine-tuned. Open-model runs are executed on an internal Slurm GPU cluster. Most 7B–8B LVLM runs use a single NVIDIA A6000 or A100 GPU. Qwen2.5-VL-32B runs use four A6000 GPUs, and Qwen2.5-VL-72B runs that are executed locally use four A100-80GB GPUs. Text-only local baselines and diagnostic inference jobs use single-GPU runs. API-model experiments are run through hosted APIs: GPT-5.4-mini through the official OpenAI API and DeepSeek-V3.2 through AI Ping[4]. Local CPU jobs are used only for request orchestration, logging, and scoring. Because no training or hyperparameter search is performed, the main computational cost comes from fixed inference over the 4,885-example visual benchmark, the 500-example challenge split, and the diagnostic subsets.

A.9 Question Template Inventory

Table 24 lists the full template inventory in the final visual benchmark and challenge split. Each template defines a relation sequence over the artwork knowledge graph. Representative natural-language realizations are provided in Appendix A.10.

A.10 Evaluation Prompt Examples

All evaluation settings ask the model to return a concise answer entity. The standard image/title settings differ only in whether the prompt contains the image, the title, or both. The evidence-exposed settings append a structured evidence block.

Standard inputs. In the image-only setting, the prompt contains the artwork image and the question. In the title-only setting, it contains the artwork title and the question. In the image+title setting, it contains the artwork image, artwork title, and question.

Flat KG. The prompt contains the standard input, the question, and an unordered list of local KG

[4] https://www.qc-ai.cn/ (Accessed: May 25th 2026)

Model	Comparison	Delta	95% CI
Qwen2.5-VL-7B	image+title vs title-only	+7.35	[6.39, 8.31]
Qwen2.5-VL-7B	image+title vs image-only	+8.43	[7.45, 9.38]
Qwen2.5-VL-7B	gold-path vs image+title	+73.53	[72.22, 74.80]
Qwen3-VL-8B	image+title vs title-only	+5.77	[4.85, 6.69]
Qwen3-VL-8B	gold-path vs image+title	+71.73	[70.44, 73.04]
Qwen3-VL-8B	flat KG vs image+title	+68.80	[67.43, 70.15]
Qwen3-VL-8B	candidate-path vs flat KG	+1.29	[0.80, 1.78]

Table 21: Selected paired comparisons. Deltas are Norm-EM percentage-point differences with paired 95% bootstrap confidence intervals. Flat KG and candidate-path are evidence-exposed diagnostics.

Category	Definition
High-prior geographic guess	The prediction is a frequent art-historical country, city, or region, but does not match the gold path.
True unrelated	The prediction is not the gold answer, not a surface alias, and not an entity directly associated with the supporting path.
Wrong relation selected	The prediction is related to the artwork or an intermediate entity, but corresponds to a different relation than the one requested.
Historical-to-modern collapse	The prediction replaces a historically distinct polity with a modern country or region.
Final-hop missing	The prediction reaches a related intermediate entity, but fails to complete the last relation needed for the answer.
Valid surface alias	The prediction is a surface variant of the gold answer under the alias diagnostic, but not an exact normalized match.
Answer type mismatch	The prediction has the wrong semantic type for the question, such as returning a country when the question asks for a capital.

Table 22: Definitions of semantic error categories used in the error analysis.

triples. The triples are serialized as subject-relation-object records. For example, the evidence block may contain:

$$(\textit{artwork}, \textit{creator}, \textit{Vincent van Gogh}),$$
$$(\textit{Vincent van Gogh}, \textit{place of birth}, \textit{Zundert}),$$
$$(\textit{Zundert}, \textit{country}, \textit{Netherlands}).$$

Other nearby triples may also be included, depending on the evidence budget.

Candidate path. The prompt contains the standard input, the question, and a list of candidate multi-hop paths. Each path is serialized as a readable sequence of concrete entity and relation labels, For example, a path may be serialized as *Wheat Field with Cypresses – creator – Vincent van Gogh – place of birth – Zundert – country – Netherlands*. A candidate path may match the gold supporting chain, while other candidates can correspond to different relation families or distractors.

Gold path. The prompt contains the standard input, the question, and only the supporting fact chain used to derive the gold answer. This is an oracle evidence setting and is not used to evaluate retrieval.

A.11 Evidence Budget and Serialization

Table 25 summarizes the evidence budgets used in the final main-set evaluation prompts. Flat-KG prompts contain unordered triples from the local graph around the linked artwork entity. Candidate-path prompts contain a bounded list of concrete multi-hop paths. Gold-path prompts contain exactly one curated supporting path.

Flat KG is built by expanding from each example's linked artwork entity over a whitelist of template relations in the frozen graph, with a cap of 100 triples. Candidate-path evidence is drawn from the precomputed candidate path pool for the same artwork entity and ranked by lexical overlap between the question and candidate template or path labels. The top candidate paths are serialized in the prompt.

Non-oracle evidence construction does not use the gold answer label, gold answer QID, gold template ID, or gold supporting-path labels to select evidence. Gold fields are retained only as evaluation metadata and for post-hoc inclusion checks. Because the local graph has high coverage, the gold supporting path is often included in the exposed evidence, with 99.96% inclusion for Flat KG and 100.00% inclusion for Candidate Path. This is why these settings are interpreted as evidence-exposure

Model	Access	Role	Notes
Qwen2.5-VL-3B/7B/32B/72B	Open model	Open LVLM, scale diagnostic	Qwen2.5-VL family; no fine-tuning.
Qwen3-VL-8B	Open model	Open LVLM	No fine-tuning.
InternVL2.5-8B	Open model	Open LVLM	No fine-tuning.
LLaVA-OneVision-7B	Open model	Open LVLM	No fine-tuning.
GPT-5.4-mini	OpenAI API	API LVLM reference	Model ID: `gpt-5.4-mini`; accessed May 2026.
Qwen2.5-7B-Instruct	Open model	Text-only LM	Title/evidence settings only.
Llama-3.1-8B-Instruct	Open model	Text-only LM	Title/evidence settings only.
DeepSeek-V3.2	AI Ping API	Text-only LM	Title/evidence settings only; accessed May 2026.

Table 23: Model identifiers and access modes. All models are evaluated under the same concise-answer protocol whenever applicable.

Template ID	Relation sequence	Main	Chal.
Main benchmark templates			
`creator_deathplace_2`	creator → place of death	577	0
`creator_birthplace_2`	creator → place of birth	570	0
`creator_movement_2`	creator → movement	563	0
`creator_citizenship_2`	creator → country of citizenship	459	0
`creator_citizenship_capital_3`	creator → country of citizenship → capital	399	0
`collection_country_capital_3`	collection → country → capital	378	0
`location_country_2`	location → country	342	0
`collection_country_2`	collection → country	329	0
`creator_birthplace_country_3`	creator → place of birth → country	278	0
`depicts_occupation_2`	depicts → occupation	277	0
`creator_deathplace_country_3`	creator → place of death → country	270	0
`depicts_citizenship_capital_3`	depicts → country of citizenship → capital	223	0
`depicts_birthplace_country_3`	depicts → place of birth → country	213	0
`admin_country_2`	located in administrative entity → country	7	0
Challenge split templates			
`collection_headquarters_2`	collection → headquarters location	0	357
`collection_headquarters_country_3`	collection → headquarters location → country	0	143

Table 24: Full template inventory in the final visual benchmark and challenge split. Counts are computed on the 4,885-example main visual benchmark and the 500-example challenge split.

Evidence setting	Min	Mean	Median	Max
Flat KG triples / prompt	2	17.29	15	100
Candidate paths / prompt	1	6.75	7	12
Candidate triples / prompt	2	16.45	17	34
Gold paths / prompt	1	1.00	1	1
Gold triples / path	2	2.36	2	3

Table 25: Evidence budget for the final ArtHopQA main-set evaluation prompts. Candidate triples count all triples serialized across the candidate paths shown to the model.

diagnostics rather than open-ended retrieval systems.

Evidence is serialized with concrete labels, e.g., *Bacchanalia (Rubens)* – creator – *Peter Paul Rubens*. The evidence shown to the model does not use QID or PID identifiers. Graph-level statistics, including the number of unique QID nodes, triples, and relation PIDs in the frozen source graph, are reported in Appendix A.4.

A.12 Supplementary Metric Diagnostics

The main paper reports Norm-EM as the primary metric. We additionally report two rule-based diagnostic metrics for surface variation and multi-occupation entities, together with an LLM-judge sensitivity analysis of Norm-EM false negatives.

Surface-alias matching. This diagnostic counts a prediction as correct if it matches either the canonical gold answer label or a documented surface alias. Alias matching is used only for diagnostic analysis and does not change the official gold answers.

Occupation-relaxed matching. This diagnostic is applied only to `depicts_occupation_2` questions. For each example, we collect P106 occupation labels of the depicted entity from the frozen

local KG. A prediction is counted as occupation-relaxed correct if it matches one of these local occupation labels. Because this uses only occupations present in the frozen graph, it is conservative and may undercount valid occupations that exist outside the local KG snapshot.

Historical entities. Historically distinct entities are not collapsed into modern ones. For example, Dutch Republic is not treated as Netherlands, and Republic of Venice is not treated as Italy.

LLM-judge sensitivity analysis. For Qwen2.5-VL-72B, the Image+Title condition contains 1,801 Norm-EM-correct predictions among 4,885 examples, leaving 3,084 Norm-EM-negative outputs. We uniformly sample 250 of these negative outputs for semantic auditing. The first-stage audit identifies 15 borderline cases. These 15 cases are mixed with 15 exact-match positive controls and 15 previously judged non-equivalent controls, then rejudged in five separately shuffled, blind GPT-5.6 Sol Pro sessions.

Twelve of the 15 borderline cases are confirmed semantically equivalent, and one of the 15 prior-negative controls is reclassified as equivalent. This gives a conservative correction rate of 13/250 = 5.20% (Wilson 95% CI: 3.06–8.69). Applying this sampled correction to the full set yields an estimated Image+Title accuracy of 40.15% (95% CI: 38.80–42.36), compared with 36.87% under Norm-EM. All 15 exact-match positive controls remain positive, and 44/45 items receive unanimous judgments across the five sessions.

The five adjudications are separate sessions of the same judge model, not independent human annotators or different model families. The remaining 220 first-stage negatives are conservatively retained as non-equivalent. We therefore use this analysis only as a sensitivity check on Norm-EM rather than as a replacement for the primary metric.