MMArt: A Multi-Perspective Multimodal Dataset for Visual Art Understanding

Shuai Wang
s.wang3@uva.nl
University of Amsterdam
Amsterdam, NL

Wangyuan Ding
w.ding@uva.nl
University of Amsterdam
Amsterdam, NL

Yixian Shen
y.shen@uva.nl
University of Amsterdam
Amsterdam, NL

Jia-Hong Huang
j.huang@uva.nl
University of Amsterdam
Amsterdam, NL

Stevan Rudinac
s.rudinac@uva.nl
University of Amsterdam
Amsterdam, NL

Monika Kackovic
m.kackovic@uva.nl
University of Amsterdam
Amsterdam, NL

Nachoem Wijnberg
n.m.wijnberg@uva.nl
University of Amsterdam
Amsterdam, NL

Marcel Worring
m.worring@uva.nl
University of Amsterdam
Amsterdam, NL

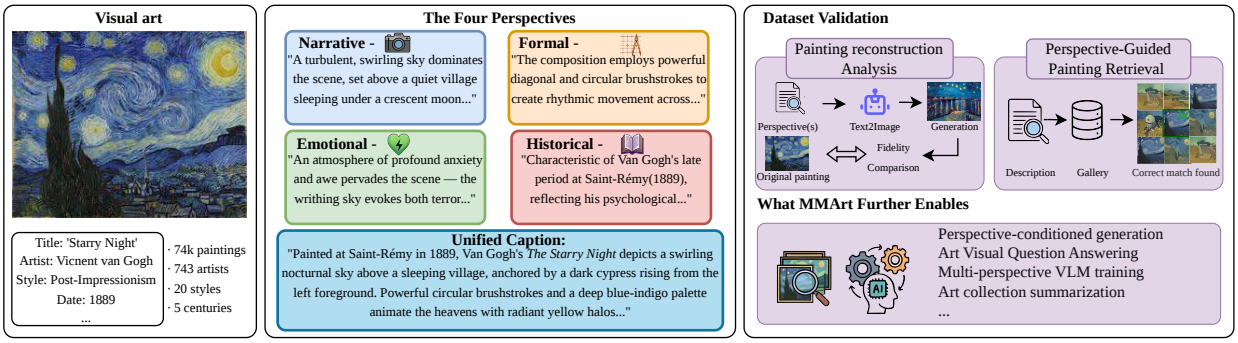

Figure 1: Overview of the MMArt dataset. Each painting is annotated with four independently generated perspectives — Narrative, Formal, Emotional, and Historical that are harmonized into a unified caption. Perspective validity is evaluated through a text-to-image reconstruction and perspective guided retrieval experiment measuring fidelity.

Abstract

Recent vision-language models demonstrate impressive general visual understanding, yet their art interpretation remains shallow: they describe surface content but struggle with formal analysis, grounded historical interpretation, or affective characterization. This is not a model but a dataset limitation. Existing art datasets are single perspective resources, where no dataset provides narrative, formal, emotional, and historical perspectives simultaneously for the same artworks. We introduce MMArt, a large-scale dataset of 74,234 WikiArt paintings, each annotated with four independently generated perspectives plus a harmonized unified caption, produced by specialized vision-language models or human annotation and validated through complementary quality evaluations. Two complementarity analyses establish that perspectives encode genuinely distinct information. A generative analysis shows that formal analysis descriptions best preserve compositional style, and historical descriptions carry strong affective signal in reconstructed images. A discriminative retrieval analysis reveals task-asymmetry: narrative descriptions drive retrieval (R@1 = 44.0%), while formal descriptions, strongest for reconstruction, are nearly non-discriminative at retrieval scale (R@1 = 7.8%). Leave-one-out analysis further confirms that historical descriptions are the least replaceable perspective across both tasks. Together, the two analyses establish that no single perspective suffices for all tasks, directly motivating MMArt's multi-perspective design. Dataset, code and supplementary materials: https://shuaiwang97.github.io/MMArt.

CCS Concepts

• **Applied computing** → **Fine arts**; • **Computing methodologies** → **Natural language generation**.

Keywords

Artwork Analysis, Multi-perspective Dataset, Multimodal Reasoning, Vision-Language Models, Art Understanding

ACM Reference Format:
Shuai Wang, Wangyuan Ding, Yixian Shen, Jia-Hong Huang, Stevan Rudinac, Monika Kackovic, Nachoem Wijnberg, and Marcel Worring. 2026. MMArt: A Multi-Perspective Multimodal Dataset for Visual Art Understanding. In *Proceedings of Proceedings of 2026 International Conference on Multimedia (MM '26).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Understanding a painting is not a simple act of perception. It is a layered process that draws simultaneously on visual attention, affective response, formal knowledge, and historical memory [4, 16, 22]. A viewer encountering Vermeer's *Girl with a Pearl Earring*[1] may

MM '26, Rio de Janeiro, Brazil
2026. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] https://en.wikipedia.org/wiki/Girl_with_a_Pearl_Earring

first register the luminous skin and direct gaze: a narrative observation. A trained eye may then notice the sfumato technique and restricted palette: a formal judgment. The same viewer may feel unease or intimacy: an emotional response. And an art historian will situate the work within the Dutch Golden Age domestic portrait tradition: a contextual interpretation. These four acts of understanding are concurrent and interacting perspectives on a single visual object.

Recent vision-language models have demonstrated impressive capabilities in general visual understanding [2, 9, 19], yet their performance on art interpretation remains shallow: they describe surface content competently but struggle to produce formal analysis, grounded historical interpretation, or affective characterization [5, 30, 33]. We argue this is not primarily a model limitation but also a data limitation. Existing computational art datasets are single-perspective resources, each capturing one interpretive dimension in isolation (Tab. 1): SemArt [12] provides iconographic and historical commentary; ArtEmis [1] covers affective response at scale; ExpArt [14] contributes expert formal analysis for a small subset; OmniArt [27] offers broad metadata coverage but no natural-language perspectives. No existing resource provides multiple interpretive dimensions simultaneously for the same artworks. As a result, models trained on these resources learn to *describe* paintings, but not to *interpret* them in the full sense.

We introduce **MMArt**, a large-scale multi-perspective dataset of 74,234 paintings designed to close this gap. Each painting is annotated with four independently generated perspectives: Narrative and Scene Interpretation, Formal Visual Analysis, Emotional Response, and Historical and Contextual Analysis, plus a harmonized unified caption. A key design principle is perspective specialization: ensuring each perspective reflects genuine domain expertise rather than a single generalist model output. MMArt uses GalleryGPT [5] for formal analysis, ArtRAG [30] for historically grounded context, and ArtEmis annotation-conditioned generation [1] for emotional response. Perspectives are collected independently to preserve interpretive diversity and harmonized into a unified caption through an explanation synthesis step.

We validate MMArt's multi-perspective design through two complementarity analyses probing perspective distinctiveness from opposite directions. A *generative complementarity analysis* tests whether each perspective recovers a distinct visual dimension: given only text, we measure which perspectives best reconstruct style, composition, and affective tone via text-to-image generation. A *discriminative complementarity analysis* evaluates each perspective(s) as a retrieval query against the full 74k gallery. Perspective contributions prove task-asymmetric: the narrative perspective drives retrieval (R@1 = 44.0%), while formal descriptions, which achieve the highest generative fidelity, are nearly non-discriminative at retrieval scale (R@1 = 7.8%). Together, they establish that no single perspective suffices for all tasks, directly motivating MMArt's multi-perspective design. MMArt is designed to support perspective-conditioned model training, affective computing on visual art, knowledge-intensive art question answering, and retrieval-augmented generation tasks that current single perspective datasets cannot support simultaneously. Our contributions are:

- **Multi-Perspective Visual Art Dataset.** We introduce MMArt, a large-scale dataset of 74,234 WikiArt paintings each annotated with four independently generated interpretive perspectives of narrative, formal, emotional, historical, via specialized human annotation or vision-language models, and a harmonized unified caption.
- **Perspective Complementarity Analyses.** We propose a two-directional evaluation framework for multi-perspective art datasets: a *generative* perspective-to-painting reconstruction and a *discriminative* perspective-guided cross-modal retrieval, jointly probing whether perspectives encode distinct and complementary visual information.
- **Public Release.** MMArt is fully open: dataset, generation and validation pipeline are publicly available to ensure reproducible research in multimedia art understanding.

2 Related Work

We discuss related work along two key relevant dimensions: existing visual art understanding datasets and methods for multi-perspective art explanation.

2.1 Visual Art Understanding Datasets

Existing datasets for computational art understanding are predominantly single-perspective resources, each typically capturing a single interpretive dimension. SemArt [12] provides Web Gallery of Art (WGA) paintings paired with museum-catalog commentaries that blend iconographic and historical content, enabling cross-modal retrieval but offering no formal, emotional, or narrative perspective. ArtEmis [1] collected 439K crowd-sourced affective utterances across WikiArt paintings — the largest coverage of emotional response to art — but contains no formal, historical, or scene-level descriptions. ExpArt [14] contributes expert formal and historical explanations for approximately 3,500 artworks, while Artpedia [26] separates visual sentences from contextual ones for 2,930 paintings. ArtCaps [20] and earlier captioning datasets [8, 24] target general or iconographic descriptions at smaller scale. OmniArt [27] offers the broadest coverage (432K artworks) but provides only structured metadata rather than natural-language perspectives. For visual question answering on art, Garcia et al. [13] and ArtQuest [7] introduce benchmarks that probe factual and semantic understanding but again operate within a single question-answering register. The common limitation of these datasets is that each captures one slice of art interpretation: studying the interplay of formal analysis and affective response typically requires combining incompatible datasets with different scales, annotation conventions, and artwork populations. MMArt addresses this gap by providing four complementary perspectives narrative, formal, emotional, and historical simultaneously for the same 74,234 paintings, with a unified caption integrating four explicitly defined perspectives. Table 1 summarizes the interpretive coverage of existing datasets.

2.2 Perspective-conditioned Art Explanation

Prior work has addressed art explanation from individual interpretive perspectives. Bai et al. [3] come closest, generating separate content, form, and context descriptions augmented with Wikipedia knowledge retrieval. However, their framework is trained on SemArt [12], whose museum-catalog commentaries blend multiple topics without dedicated per-perspective annotation — with more than 80% of paintings missing at least one perspective. GalleryGPT [5]

Table 1: Comparison of art understanding datasets across interpretive dimensions: Narr. = Narrative, Form. = Formal Analysis, Emot. = Emotion, Hist. = Historical Context. ✓ = dedicated annotation; ∼ = partial or metadata only; – = absent.

Dataset	Size	Narr.	Form.	Emot.	Hist.	Unified
SemArt [12]	21K	✓	∼	–	✓	–
ArtEmis [1]	80K	–	–	✓	–	–
Artpedia [26]	3K	✓	–	–	∼	–
ExpArt [14]	3.5K	–	✓	–	✓	–
ArtCaps [20]	4K	✓	–	–	–	–
OmniArt [27]	432K	–	–	–	∼	–
MMArt (ours)	**74K**	✓	✓	✓	✓	✓

fine-tunes a vision-language model for formal compositional analysis on the PaintingForm dataset; ArtGPT-4 [33] targets general artistic understanding via an LLaVA adapter. Both confirm that specialized models outperform generalists on individual perspectives, yet neither combines multiple perspectives into a unified dataset. Knowledge grounding is equally critical for historical interpretation, where pure visual inference leads to frequent factual confabulation [15]. ArtRAG and VL-KGE [10, 30] demonstrates that art-historical knowledge graphs substantially improve factual accuracy and interpretive depth. The key departure of MMArt from all prior work is scale and coverage: where existing resources capture one interpretive dimension independently, MMArt provides all four simultaneously for 74k paintings, enabling the cross-perspective complementarity analyses that form the core of our validation.

3 MMArt Dataset Construction

The four-perspective architecture is grounded in Panofsky's framework [22] for layered art interpretation, which distinguishes factual description, formal analysis, and symbolic meaning as distinct and non-reducible acts of understanding. We operationalize this as four computational perspectives: narrative description, which supports scene-level retrieval and VQA; formal analysis, which enables style transfer and compositional modeling; emotional response, which grounds affective computing research; and historical context, which supports knowledge-intensive generation and RAG benchmarks. Each perspective targets a distinct downstream task family that no existing single-perspective dataset can support simultaneously — this is the core design rationale for MMArt. MMArt is built on WikiArt[2], a publicly accessible repository that has served as the foundation for numerous computational art understanding benchmarks [1, 11, 31]. It covers approximately 75k artworks across 20 style categories spanning the 15th through the 21st century, each associated with structured metadata including title, artist name, production date, style, and school. The remainder of this section describes the four-perspective architecture, generation pipeline, and quality control procedures.

3.1 Four-Perspective Architecture

Rather than generating a single descriptive caption, MMArt decomposes each artwork's annotation into four independently generated perspectives and one unified description:

$$P(i) = \{e^{\text{narr}},\ e^{\text{form}},\ e^{\text{emot}},\ e^{\text{hist}}\ e^{\text{unif}}\}, \tag{1}$$

where i denotes a painting and each component captures a distinct interpretive dimension. *Narrative and Scene Interpretation (e^{narr}):* A factual account of the depicted entities, figures, scene composition, and visual elements as they appear in the image. This perspective answers *what* is shown, without invoking symbolic or historical interpretation. *Formal Visual Analysis (e^{form}):* An analysis of the painting's compositional structure, including spatial organization, color palette, brushwork, use of light and shadow, and visual rhythm. This perspective corresponds to the vocabulary of formal art criticism, focusing on *how* the work is constructed. *Emotional Response (e^{emot}):* An affective characterization of the painting's perceived mood, atmosphere, and psychological tone by a viewer. This perspective captures the phenomenological dimension of art encounter—*what it feels like* to look at the work. *Historical and Contextual Analysis (e^{hist}):* An art-historical interpretation situating the work within its cultural context: movement affiliation, iconographic codes, period-specific symbolic meaning, and relevant biographical or historical circumstances. This perspective answers *why* the work looks and means as it does.

In addition to the four independent perspectives, each painting receives a *unified caption* e^{unif} produced by harmonizing all four into a single coherent description. This unified caption serves as a strong single-text baseline in benchmark evaluations.

3.2 Generation Pipeline

A central design principle of MMArt is perspective specialization: each perspective is generated by a model chosen for its demonstrated strength in the corresponding interpretive task. This is a deliberate data construction strategy, not a fragmentation of interpretation. Generating perspectives independently with specialized models ensures each dimension reflects genuine domain expertise, providing the clean per-perspective supervision that joint multi-perspective VLM training requires.

Narrative perspective. Qwen3-VL-8B-Instruct [2] is selected for its broad visual pretraining that minimizes art-domain bias, keeping output anchored in observable scene content. Narrative captions are generated from the painting image and metadata alone using a prompt (π_{narr}) that explicitly prohibits symbolic, historical, or emotional reference, constraining output to visually observable entities, figures, and spatial relationships.

Formal analysis. Formal perspectives are generated using specialized GalleryGPT [5], a LLaVA-7B [19] model fine-tuned on its PaintingForm dataset specifically for formal art analysis. No general-purpose VLM matches GalleryGPT's adherence to formal criticism vocabulary on this task. Its supervised fine-tuning on expert-annotated PaintingForm data makes it the only publicly available model purpose-built for this perspective. The prompt (π_{form})instructs the model to analyze compositional structure, palette, brushwork, and spatial organization while suppressing narrative or interpretive content.

Emotional response. Emotional perspectives are generated using Qwen3-VL-8B-Instruct [2], conditioned on the painting image together with human-written affective utterances from ArtEmis [1,

[2] https://www.wikiart.org

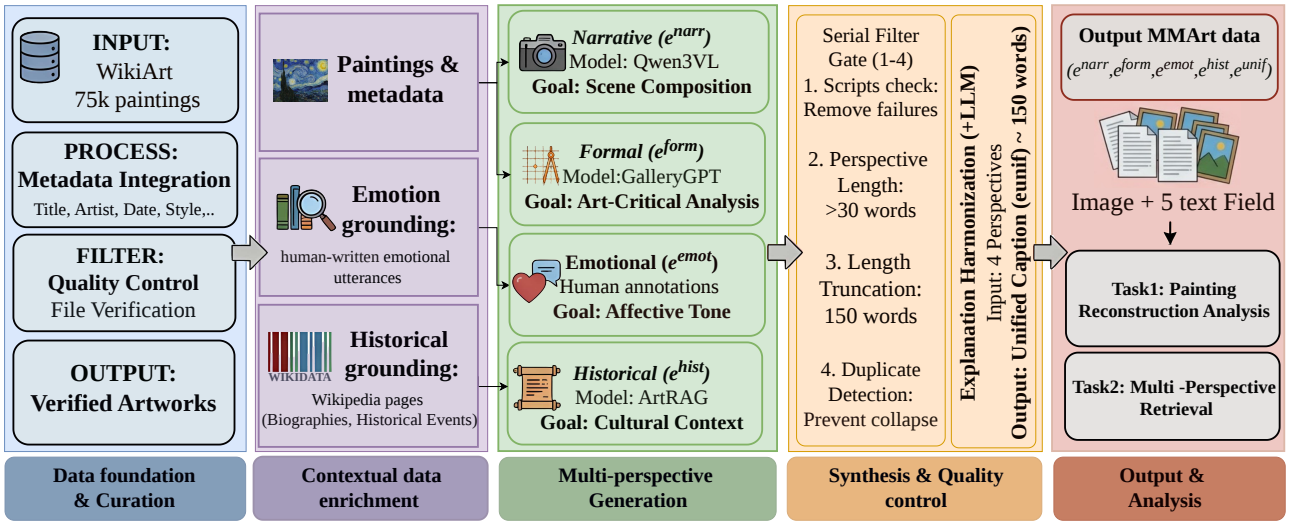

Figure 2: Overview of the MMArt dataset construction pipeline. Each painting is processed through four specialized vision-language models to produce independently generated perspectives (narrative, formal, emotional, historical), which are then harmonized into a unified caption.

21], a large-scale emotion dataset collected over the same WikiArt corpus, providing on average 5.7 authentic human responses per painting. This overlap ensures that affective grounding draws on real viewer reactions to the exact paintings in MMArt rather than out-of-domain sentiment signals. Qwen3-VL is chosen here because its instruction-following capability allows reliable conditioning on the ArtEmis utterances as affective anchors. The prompt (π_{emot}) instructs the model to characterize perceived mood and psychological tone grounded in the provided human utterances, without describing scene or historical context.

Historical and contextual analysis. Historical perspectives are generated using ArtRAG [30] augmented with structured art-historical context retrieved via . For each artwork, the top-5 context documents covering artist biography, movement affiliation, and relevant historical events are retrieved by embedding ranking and concatenated as generation context. This retrieval-augmented approach is chosen over direct VLM generation because historical and biographical facts are not recoverable from visual appearance alone [30]. Instead, grounding generation in verified external knowledge substantially reduces factual error. The prompt (π_{hist}) instructs the model to situate the work within its cultural and art-historical context using the retrieved documents, explicitly suppressing scene description and subjective affect.

Unified caption. Given the four independently generated perspectives $\{e^{\text{narr}}, e^{\text{form}}, e^{\text{emot}}, e^{\text{hist}}\}$, a unified caption e^{unif} is produced via Qwen3-8B [2] using a harmonization prompt that synthesizes the four perspectives into a single coherent, non-redundant description of approximately 150 words. The unified caption preserves interpretive breadth while eliminating cross-perspective repetition, and is designed for downstream tasks such as retrieval and language-model grounding where a holistic single-text representation is preferred. Formally, the generation of perspective v for painting w with image I, metadata M, and optional retrieved context $\mathcal{S}$ is:

$$e^{v} = \text{VLM}_{v}(I,\ M,\ \mathcal{S},\ \pi_{v})\,, \tag{2}$$

where π_v is a perspective-specific prompt template designed to enforce interpretive focus and minimize cross-perspective overlap. All perspectives are generated with model's default decoding temperature and max_tokens = 256. Full prompt templates for all five generation steps are provided in the supplementary website.

3.3 Quality Control and Data Characterization

Text quality filtering. We apply four automated cleaning steps to the raw generated output. (i) *Script contamination*: perspectives containing non-Latin characters are filtered out, as these indicate generation failure in which the model switches output language rather than describing the painting. (ii) *Length filtering*: e length under 30 words are nulled and filtered out as uninformative, and explanations exceeding 150 words is truncated at the nearest sentence boundary to remove generation runoff without discarding content. (iii) *Duplicate detection*: exact-duplicate descriptions values across paintings are nulled, as these indicate model collapse on visually similar inputs. The full dataset retains all 75,336 paintings with nulls preserved; the experiment-ready subset requires all four perspectives to be non-null, yielding 74,234 paintings.

A core claim of MMArt is that the four perspectives encode distinct information rather than stylistic paraphrases of each other. We validate this at two levels.

Semantic distinctiveness. To quantify the degree of overlap between perspectives for the same painting, we compute pairwise cosine similarity between perspective embeddings using CLIP ViT-L/14 text encodings on a random sample of 1,000 paintings. Table 2 reports the mean pairwise similarity between all perspective pairs. All pairs exhibit low cosine similarity (<0.55), with the narrative–formal pair showing the greatest divergence and narrative–historical showing the most overlap, consistent with

the shared factual grounding of both perspectives. These results confirm that the perspectives are semantically non-redundant.

Table 2: Mean pairwise CLIP-text cosine similarity across perspectives. Lower indicate greater semantic distinctiveness.

	e^{narr}	e^{form}	e^{emot}	e^{hist}
e^{narr}	1.00	0.41	0.48	0.52
e^{form}	0.41	1.00	0.43	0.44
e^{emot}	0.48	0.43	1.00	0.46
e^{hist}	0.52	0.44	0.46	1.00

Table 3: LLM-as-judge quality ratings (1–5 scale, mean±std.)

Perspective	Fidelity	Accuracy	Depth
Narrative	4.49±0.80	4.33±0.95	3.77±0.45
Formal	3.70±1.17	3.62±1.32	3.03±0.74
Emotional	4.99±0.12	4.41±0.71	3.98±0.17
Historical	4.86±0.42	4.28±0.76	3.78±0.43
Unified	4.64±0.68	4.13±1.04	4.08±0.43

LLM-as-judge evaluation. We evaluate all five perspectives using Gemma-3-27B [28] as an independent vision-language judge on 300 randomly stratified paintings. As a model from a different family than all generation models used in dataset construction, Gemma-3-27B mitigates self-evaluation bias. Each perspective is rated on three dimensions (1–5 scale): *perspective fidelity* (if the text focuses on the correct aspect), *factual accuracy* (if the content is correct, assessed against both the painting image and metadata), and *depth* (if it provides substantive detail beyond generic description). Each perspective is rated independently in a single zero-shot pass; the rating prompt instructs the judge to assess only the specified dimension without reference to other perspectives. The full judge prompt is provided in the supplementary materials. Table 3 reports the results. Three patterns are notable. First, e^{emot} achieves near-perfect fidelity (4.99±0.12) with the lowest variance of any perspective, reflecting the accurate information of ArtEmis affective labels. Second, e^{form} is the weakest perspective across all three dimensions, particularly in depth (3.03±0.74), attributable to GalleryGPT's base model (LLaVA-7B) smaller model capacity relative to the models used for narrative and historical generation. Third, e^{unified} achieves the highest depth score (4.08±0.43) of any perspective, confirming that the harmonization step produces captions with richer substantive detail than any single-perspective source. Figure. 3 reports the distribution of constructed MMArt dataset.

Ethical Considerations and License. MMArt is built entirely from WikiArt, a publicly accessible repository of artworks whose copyright has expired or which have been made available for educational use. The dataset contains no personal data, private images, or crowd-sourced human annotations beyond the published ArtEmis dataset [1], which was collected under institutional review. All generated perspectives describe artworks, not individuals. No licence terms of the source materials are violated. MMArt is released under the Creative Commons Attribution 4.0 International license (CC BY 4.0) for academic and research use.

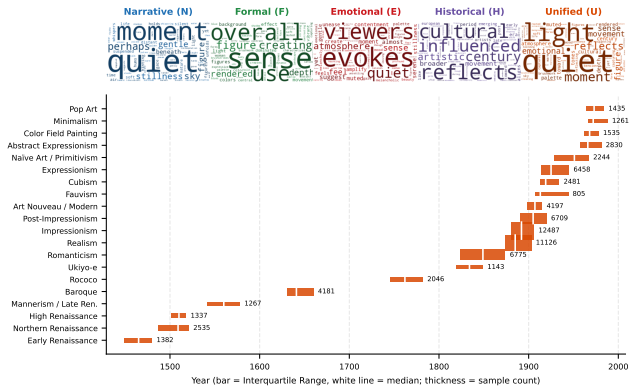

Figure 3: Top: Characteristic vocabularies for different perspectives, using TF-IDF weighted word clouds. Bottom: Chronological distribution and volume of art styles.

4 Perspective Complementarity Analyses

Having established that the four perspectives are semantically non-redundant in Section 3.3, a deeper question remains: Do they contribute differently across tasks? A perspective may encode distinct information yet still be uninformative for a particular downstream application. We introduce a two-directional analysis framework to probe precisely this: whether each perspective's utility is task-asymmetric. Both analyses evaluate nine conditions: four singles (e^{narr}, e^{form}, e^{emot}, e^{hist}), four leave-one-out triples (each omitting one perspective), and the unified caption e^{unif}. Conditions are evaluated on the same stratified 1,000 painting samples.

4.1 Two-directional Framework

We designed two-directional analysis framework for the experiment. The generative analysis investigates whether a perspective can *reconstruct* a painting, which probing what visual information each perspective encodes. Given text from one or more perspectives without the original image, a text-to-image model generates a reconstruction of the original painting. Multi-perspective conditions are first synthesized into a single coherent ∼80-word prompt using Qwen3-8B, ensuring all conditions enter the image generator in the same text format so that differences in fidelity are attributable to perspective content rather than formatting. Two text-to-image models including FLUX.2-Klein-4B [6] and Qwen-Image-2512 [32] are used to confirm that findings are model-agnostic. Reconstruction fidelity is measured against the original painting across three complementary metrics:

$$\mathcal{F}(w, \hat{w}) = \left(\delta_{\text{style}},\ \delta_{\text{comp}},\ \delta_{\text{emot}}\right), \tag{3}$$

where δ_{style} is CLIP ViT-L/14 [23] cosine similarity and δ_{comp} is DINOv3-vitl16-pretrain-lvd1689m embedding cosine similarity. The two metrics are complementary by design: CLIP's image–text contrastive training makes it sensitive to global style and semantic content, while DINOv3's self-supervised image-only training yields representations more attuned to local spatial structure and compositional layout [25]. Together they capture different visual dimensions of reconstruction fidelity that a single metric would conflate [29]. δ_{emot} is CLIP zero-shot emotion agreement, measuring the proportion of paintings for which the regenerated image and the original share the same top-1 ArtEmis emotion label.

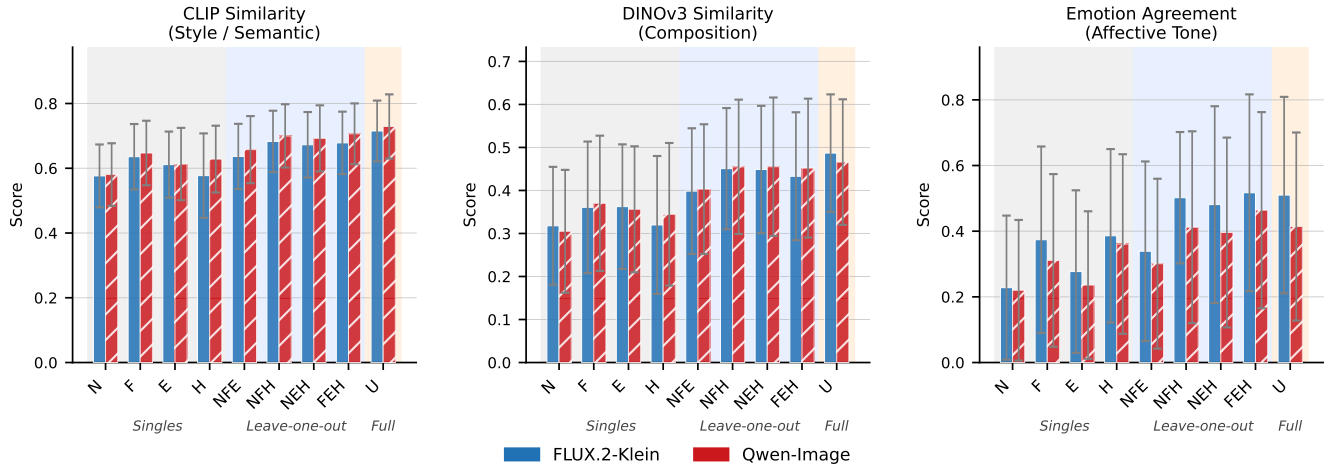

Figure 4: Reconstruction fidelity across nine perspective conditions for FLUX.2-Klein (blue) and Qwen-Image (red). Shaded regions separate singles, leave-one-out triples, and the full four-perspective condition. Error bars show ±1 standard deviation.

The discriminative direction asks whether a perspective can *identify* a painting among all MMArt candidates, probing how unique the encoded information is. A perspective that reconstructs well may still fail to discriminate, and vice versa; only together do the two directions establish genuine complementarity. We embed all 74,234 MMArt paintings as a fixed gallery using Qwen3-VL-Embedding-2B [18] and Jina-CLIP-v2 [17], two architecturally distinct embedding models, to confirm that findings are model-agnostic. Given a textual perspective as a query, we evaluate description to painting retrieval on the same stratified 1,000-painting sample and nine perspective conditions. Retrieval performance is measured by Recall at k (R@k for $k \in \{1, 5, 10\}$), Precision at k (P@k for $k \in \{1, 5, 10\}$) and NDCG, Mean Reciprocal Rank (MRR), all computed over the full 74k painting gallery.

Painting Reconstruction Analysis Results. Figure 4 reports reconstruction fidelity across all nine conditions for image generators. The unified perspective condition e^{unif} outperforms all single perspectives on both CLIP style similarity and DINOv3 composition, confirming that combining perspectives yields richer visual grounding than any individual caption. Among singles, e^{form} achieves the highest CLIP fidelity, while FEH leads on emotion agreement, indicating that e^{narr} contributes less to affective tone than the remaining perspectives. Leave-one-out analysis confirms that e^{hist} is the hardest perspective to replace, contributing the largest marginal gain across all three metrics. Together, these results show that each perspective recovers a distinct subset of fidelity dimensions, and reconstruction improves as perspectives are combined.

Perspective-Guided Painting Retrieval Results Figure 5 reports retrieval performance across all nine conditions. e^{narr} is the strongest single perspective (R@1 = 44.0%, MRR = 0.537), while e^{hist} is nearly non-functional as a retrieval query (R@1 = 2.2%, MRR = 0.037). This ordering is the inverse of the regeneration results in Figure 4, where e^{form} achieves the highest CLIP fidelity and e^{narr} the lowest. Leave-one-out analysis further reveals that e^{hist} is the least replaceable perspective across all three fidelity metrics, whilee^{narr}'s contribution to style and composition is partially redundant when the other three perspectives are present. The unified caption e^{unified} (R@1 = 45.1%, MRR = 0.548) marginally outperforms e^{narr} by incorporating complementary scene detail from all four perspectives.

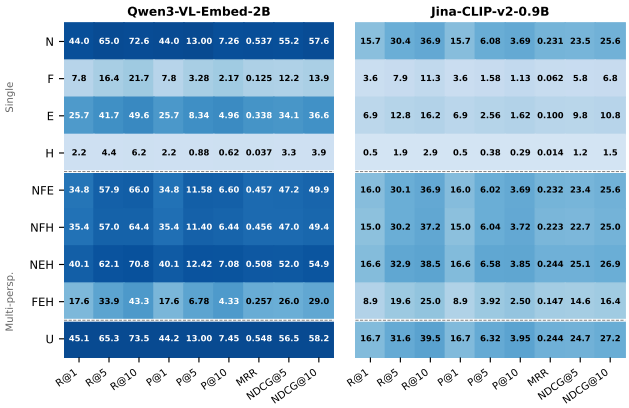

Figure 5: Description to painting retrieval performance across perspective conditions from the full MMArt gallery.

5 Conclusion

We presented MMArt, a large-scale dataset of 74,234 WikiArt paintings, each annotated with four independently generated interpretive perspectives and a harmonized unified caption. Two complementarity analyses establish that perspective utility is task-asymmetric. No single perspective suffices across reconstruction, retrieval, and generation, and this directly shows the effectiveness of MMArt's multi-perspective design. Leave-one-out analysis shows that historical descriptions are the least replaceable perspective across style, composition, and affective tone. The MMArt further enable perspective-conditioned VLM training, affective computing research, and knowledge-intensive art question answering and retrieval-augmented generation. Dataset, generation scripts, and pre-computed embeddings are publicly released to support reproducible research across all these applications.

References

[1] 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)*.

[2] 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, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-VL Technical Report. arXiv:2511.21631 [cs.CV] https://arxiv.org/abs/2511.21631

[3] Zechen Bai, Yuta Nakashima, and Noa García. 2021. Explain Me the Painting: Multi-Topic Knowledgeable Art Description Generation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 5402–5412. https://api.semanticscholar.org/CorpusID:237490413

[4] Michael Baxandall. 1988. *Painting and experience in fifteenth century Italy: a primer in the social history of pictorial style*. Oxford Paperbacks.

[5] Yi Bin, Wenhao Shi, Yujuan Ding, Zhiqiang Hu, Zheng Wang, Yang Yang, See-Kiong Ng, and Heng Tao Shen. 2024. GalleryGPT: Analyzing Paintings with Large Multimodal Models. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

[6] Black Forest Labs. 2025. FLUX.2: Analyzing and Enhancing the Latent Space of FLUX – Representation Comparison. https://bfl.ai/research/representation-comparison.

[7] Tibor Bleidt, Sedigheh Eslami, and Gerard De Melo. 2024. ArtQuest: Countering Hidden Language Biases in ArtVQA. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI, USA.

[8] Eva Cetinic. 2021. Iconographic Image Captioning for Artworks. In *Pattern Recognition. ICPR International Workshops and Challenges*, Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani (Eds.).

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*.

[10] Athanasios Efthymiou, Stevan Rudinac, Monika Kackovic, Nachoem Wijnberg, and Marcel Worring. 2026. VL-KGE: Vision-Language Models Meet Knowledge Graph Embeddings. *arXiv preprint arXiv:2603.02435* (2026).

[11] Athanasios Efthymiou, Stevan Rudinac, Monika Kackovic, Marcel Worring, and Nachoem Wijnberg. 2021. Graph Neural Networks for Knowledge Enhanced Visual Representation of Paintings. In *Proceedings of the 29th ACM International Conference on Multimedia*.

[12] Noa Garcia and George Vogiatzis. 2018. How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In *Proceedings of the European Conference in Computer Vision Workshops*.

[13] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. 2020. A Dataset and Baselines for Visual Question Answering on Art. http://arxiv.org/abs/2008.12520

[14] Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. [n. d.]. Towards Artwork Explanation in Large-scale Vision Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, Article 248 (March 2023), 38 pages.

[16] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. 2014. Recognizing Image Style. In *Proceedings of the British Machine Vision Conference*. BMVA Press.

[17] Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. 2025. jina-clip-v2: Multilingual Multimodal Embeddings for Text and Images. arXiv:2412.08802 [cs.CL] https://arxiv.org/abs/2412.08802

[18] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking. arXiv:2601.04720 [cs.CL] https://arxiv.org/abs/2601.04720

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, Vol. 36.

[20] Yue Lu, Chao Guo, Xingyuan Dai, and Fei-Yue Wang. 2024. ArtCap: A Dataset for Image Captioning of Fine Art Paintings. *IEEE Transactions on Computational Social Systems* (2024).

[21] Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, and Mohamed Elhoseiny. 2022. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 21263–21272.

[22] E. Panofsky. 1955. *Meaning in the Visual Arts*. University of Chicago Press. https://books.google.nl/books?id=Qsa00QEACAAJ

[23] 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. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[24] Shurong Sheng and Marie-Francine Moens. 2019. Generating Captions for Images of Ancient Artworks. In *Proceedings of the 27th ACM International Conference on Multimedia*.

[25] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. 2025. DINOv3. arXiv:2508.10104 [cs.CV] https://arxiv.org/abs/2508.10104

[26] 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*.

[27] Gjorgji Strezoski and Marcel Worring. 2018. OmniArt: A Large-scale Artistic Benchmark. *ACM Trans. Multimedia Comput. Commun. Appl.* (2018).

[28] Gemma Team. 2025. Gemma 3 Technical Report. arXiv:2503.19786 [cs.CL] https://arxiv.org/abs/2503.19786

[29] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9568–9578.

[30] Shuai Wang, Ivona Najdenkoska, Hongyi Zhu, Stevan Rudinac, Monika Kackovic, Nachoem Wijnberg, and Marcel Worring. 2025. ArtRAG: Retrieval-Augmented Generation with Structured Context for Visual Art Understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*.

[31] Shuai Wang, Jiayi Shen, Athanasios Efthymiou, Stevan Rudinac, Monika Kackovic, Nachoem Wijnberg, and Marcel Worring. 2024. Prototype-Enhanced Hypergraph Learning for Heterogeneous Information Networks. In *MultiMedia Modeling*.

[32] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. 2025. Qwen-Image Technical Report. https://arxiv.org/abs/2508.02324

[33] Zheng Yuan, HU Xue, Xinyi Wang, Yongming Liu, Zhuanzhe Zhao, and Kun Wang. 2024. ArtGPT-4: Towards Artistic-understanding Large Vision-Language Models with Enhanced Adapter. In *Proceedings of conference on language modelling*.