Harnessing Multimodal LLMs for Attribute Specific Image Retrieval

Anonymous ACL submission

Abstract

Despite extensive research on visual style understanding, accurately capturing and comparing artistic style remains challenging. Traditional retrieval methods rely on global embeddings that conflate multiple stylistic dimensions—such as brushstrokes, color palette, lighting, and composition—into single vectors, limiting both interpretability and fine-grained retrieval. We present a training-free framework that leverages Multimodal Large Language Models (MLLMs) to decompose images into multiple attribute-specific embeddings, providing disentangled style representation. Independent representations for each stylistic attribute are obtained by extracting embeddings from intermediate transformer layers, conditioned on carefully designed user and system prompts that are generated dynamically. We demonstrate that intermediate MLLM layers possess richer visual information than the final-layer, which tends to suppress visual features to generate textual outputs. To enhance representational quality, we introduce *dynamic prefixing*, an automated approach that generates task-adaptive system and user prompts that outperform manual prompt design. For retrieval, we propose a layer-wise hard-voting fusion mechanism that aggregates evidence across multiple transformer layers without learnable parameters. Extensive experiments on WikiArt and DomainNet demonstrate that our training-free approach achieves competitive or superior performance compared to both generic vision encoders and style-specific models like CSD (Somepalli et al., 2024), while providing attribute-level interpretability. Human preference studies further validate that our attribute-based retrieval aligns more closely with human perception than global representations, being preferred 54.95% of the time over images retrieved by contemporary methods.

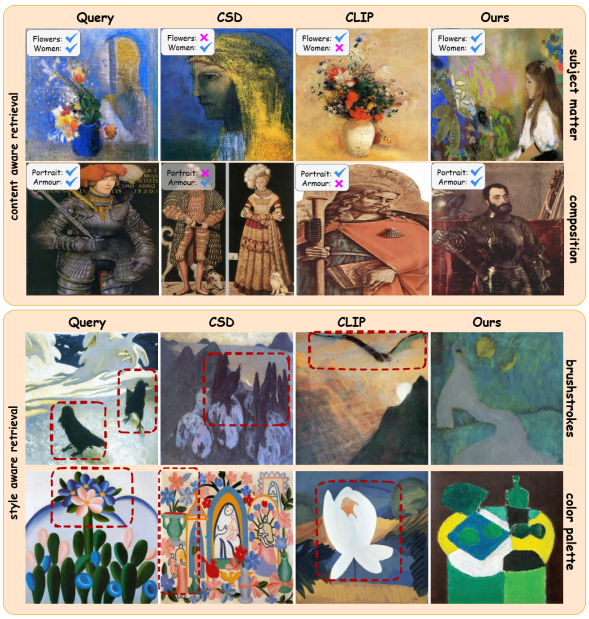

Figure 1: Previous methods primarily retrieve images using global representations. In contrast, our attribute-based retrieval, when compared with CSD (Somepalli et al., 2024) and CLIP (Radford et al., 2021), exhibits two key advantages. (Top) It enables accurate content-based retrieval by correctly identifying and retrieving images that contain the primary subjects present in the query image. (Bottom) It effectively disentangles style from content. While existing methods are often biased toward dominant objects in the image (highlighted by red boxes), our approach retrieves images according to the desired stylistic attributes.

1 Introduction

Understanding and comparing visual style in images remains a challenging problem despite extensive prior research (Tenenbaum and Freeman, 1996; Graham et al., 2012; Hughes et al., 2010; Lawrence-Lightfoot and Davis, 2002; Li et al., 2011; Sablatnig et al., 1998; Saleh and Elgammal, 2015; Gairola et al., 2020; Matsuo and Yanai, 2016; Ruta et al., 2021; Wynen et al., 2018). Many existing retrieval methods (Radford et al., 2021; Caron et al.,

2021; Pizzi et al., 2022) rely on global embeddings that conflate multiple stylistic dimensions—such as brushstrokes, color palette, lighting, and composition—into a single vector. This entanglement limits both interpretability and downstream performance, often prioritizing content over style (Somepalli et al., 2024). For instance, two paintings may share similar composition but differ significantly in color palette and brushwork (Figure 2). A single embedding averages over these dimensions, producing potentially misleading similarity scores with no mechanism to isolate specific stylistic factors. Modern diffusion-based image generators (Rombach et al., 2022; Ramesh et al., 2021; Betker et al., 2023; Pernias et al., 2023; Adobe, 2023; Midjourney, n.d.), such as Stable Diffusion (Rombach et al., 2022), learn artistic styles from large web-scale datasets (Schuhmann et al., 2022). Evaluating whether generated outputs align with training distributions or reproduce style-specific signatures requires *attribute-level* feedback, which traditional global embeddings cannot provide. A natural solution is to decompose artworks into multiple components and compare images using attribute-specific embeddings. However, training separate models for each attribute is costly and does not easily adapt to new tasks or attributes.

Based on prior work (Li et al., 2011; Karayev et al., 2014; Srinivasa Desikan et al., 2022; Somepalli et al., 2024), we first identify six attributes (Table 1) that jointly describe the style and content of an image. To extract attribute-specific embeddings, we leverage the intrinsic visual-language understanding of Multimodal Large Language Models (MLLMs). MLLMs offers several advantages: flexible attribute management without retraining, generalization across domains, and interpretable embeddings through text outputs. Unlike an image encoder such as ViT, an MLLM can be "steered" via text prompts to focus on specific features (e.g., "Describe the brushstrokes"). These attribute-level descriptions are encoded as separate embeddings, forming a **structured, disentangled representation** of visual style.

To enhance *representational* capabilities, we explore multiple strategies. First, we propose to use intermediate MLLM layers which yield superior retrieval performance compared to final-layer outputs, as hidden states preserve fine-grained visual information often compared to final textual predictions (See Figure 6). Second, we propose a dynamic prefixing routine to automatically generate system prompts and enrich MLLM hidden states, outperforming manual prompt design. Third, a layer-wise hard-voting fusion mechanism aggregates retrieval evidence across multiple transformer layers without learnable parameters, producing stable rankings that utilizes features from shallow and deeper layers. Figure 1 illustrates the advantages of our approach. Unlike traditional retrieval systems like CSD (Somepalli et al., 2024) and CLIP (Radford et al., 2021), which conflate style and content, our method separates these factors. Content-focused queries retrieve semantically related images with stylistic variation, while style-focused retrieval identifies images with similar brushstrokes, palettes, or composition regardless of subject matter. Our contributions are:

- A training-free framework that disentangles attribute-level style information (color, brushstrokes, composition, lighting, abstraction, subject matter), to represent each image with multiple attribute-specific embeddings for fine-grained and interpretable image retrieval.
- Analysis showing that the intermediate layers of MLLMs embed richer visual representations than final-layer outputs, with systematic evaluation of embeddings extracted across intermediate MLLM transformer layers.
- Investigation of prompting strategies and proposing a *dynamic prefixing* method, which significantly outperforms fixed manual prompts for each attribute.

2 Motivation

Figure 2 illustrates two paintings with distinct visual style but shared semantic content. This raises a fundamental question in style analysis: *Are these two artworks similar or different?*

Representing style in images. Style is generally described as a distinctive manner that permits grouping of works into related categories (Fernie, 1995; Menis-Mastromichalakis et al., 2020), encompassing color, brushstrokes, composition, and perspective. Existing approaches (Somepalli et al., 2024; Lecoutre et al., 2017; Kumar et al., 2025) represent artworks using a single embedding vector. While effective for coarse similarity, such embeddings lack interpretability and fail to isolate individual stylistic factors. Consequently, they cannot indicate which attributes are aligned or divergent,

Figure 2: (Left) The Rising of the Sun by François Boucher (1753) and (Right) Lucifer and Abdiel by Gustave Doré (1858). While visually distinct at first glance, both paintings depict human figures in surreal, floating environments. Our method correctly identifies these images as most similar in subject matter while recognizing them as dissimilar in color palette across the style attributes. (Top) The MLLM responses for the images corresponding to the attributes, effectively capture these stylistic similarities (in green) and differences (in red).

limiting fine-grained control. For instance, two images may match in color tones but diverge in subject matter or abstraction. Similarly, retrieval based on specific aspects, such as brushstrokes, is not possible with a single embedding. Decomposing artworks into attribute-specific embeddings addresses these limitations. It allows precise similarity assessment along individual dimensions, supports flexible retrieval and clustering based on stylistic features, and aligns with human perceptual judgments, which consider multiple axes of style. Table 1 describes the six visual attributes used in this work. Brushstrokes encode local textural patterns and the artist's mark-making style; Color Palette reflects the global chromatic structure that shapes mood and aesthetic tone; Subject Matter provides semantic grounding distinguishing thematic domains; Composition describes the spatial arrangement and structural geometry of the scene; Lighting governs atmosphere, contrast, and perceptual depth; and Abstraction characterizes the level of representational realism. Collectively, these attributes represent the essential elements that characterize an image's stylistic identity.

MLLMs for open-set retrieval. Attribute-based

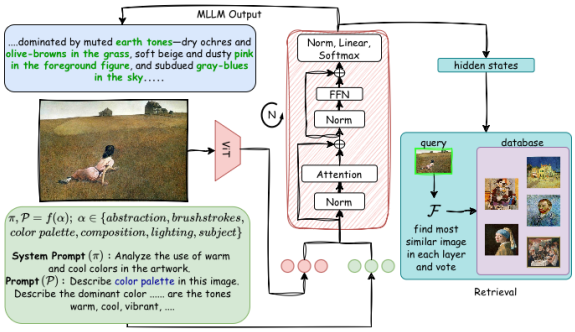

Figure 3: We leverage pretrained MLLMs for attribute-based fine-grained image retrieval. Given an image I and a target attribute α, we construct a task-specific prefix π and prompt P, which are jointly provided to the model. Hidden states from each of the N individual transformer layers are extracted and treated as layer-specific image representations. For retrieval, similarity is computed between each layer's representation and those of the query images, and a voting operator $\mathcal{F}$ aggregates the layer-wise similarities to produce the final retrieval result.

representations are powerful but inflexible when models must be retrained for each attribute or domain. MLLMs, trained on diverse datasets, provide generalizable visual understanding and enable querying for new attributes without retraining. The central research question is thus: ***How can we leverage the zero-shot understanding of MLLMs for open-set retrieval?***. Prior work has explored MLLM representations in various ways. (Zhuang et al., 2024) uses the last-layer embedding for dense document representations, while (Sun et al., 2025) averages last-layer embeddings to reveal word–image region correspondences. (Kawarada et al., 2025) prompts an MLLM to describe an image with a single condition-specific word, using the last token's hidden state as a conditional image embedding. (Liu et al., 2025) queries visual tokens in the KV-cache of transformer heads for correspondence tasks. Despite these efforts, a systematic study in effectiveness of hidden states for open-set retrieval remains largely unexplored.

3 Methodology

Our pipeline (Figure 3) consists of three main components: 1) prompt selection, 2) embedding extraction from hidden states, and 3) embedding selection or fusion for retrieval. A key advantage of our approach is relative comparison: unlike single-vector representations that output a single score with limited interpretability, our attribute-wise embedding

Table 1: Descriptions of visual attributes used for fine-grained style interpretation. For each attribute, we prompt the MLLM using the template "Describe <attribute> in this image. Context: <description>," where the contextual description specifies the visual characteristics to attend to for that attribute.

Attribute	Description / Context
Brushstrokes	Texture and style of paint application: visible, expressive, smooth, dotted, or swirling strokes.
Color Palette	Dominant colors and harmony: warm, cool, vibrant, muted, monochrome, complementary, or analogous.
Subject Matter	Main theme: portrait, landscape, still life, abstract, religious, or narrative scene.
Composition	Arrangement of elements: symmetry, focal points, dynamic flow, crowded or spacious layout.
Lighting	Light source and effect: soft, dramatic, shadows, sunlight, artificial light, or glowing highlights.
Abstraction	Level of realism: photorealistic, stylized, semi-abstract, or non-representational.

allows meaningful comparisons between images along each visual dimension, supporting more nuanced analysis and retrieval. Moreover, our method is training-free, i.e, it uses MLLM's zero-shot capabilities to retrieve attribute-specific similar images.

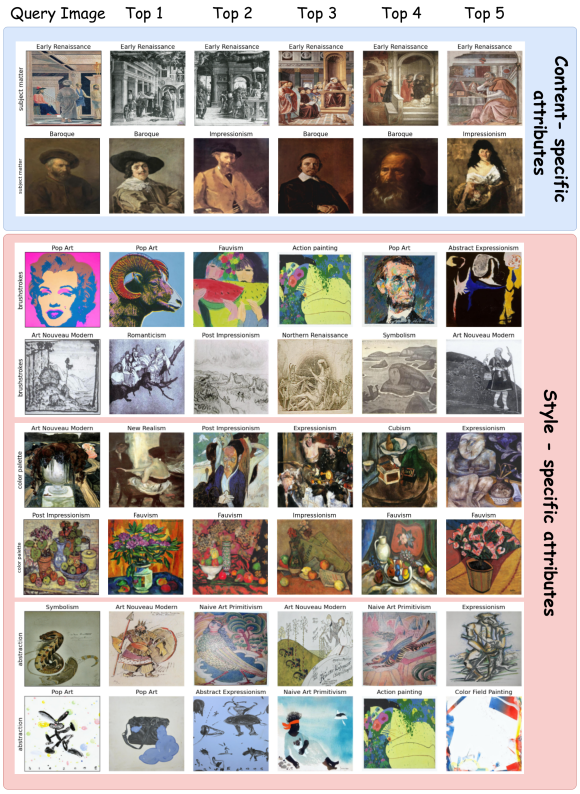

Figure 4: Top 5 images retrieved by our method for each attribute. Our approach effectively disentangles attributes from objects, enabling accurate retrieval of matching images.

3.1 Prompt Selection

Prompt design, or "priming", has been shown to significantly influence the performance of frozen LLMs (Wei et al., 2022; Long, 2023; Zhang et al., 2023), particularly in tasks requiring complex reasoning and understanding. By providing additional information during response generation, prompts can guide the model's behavior. Consequently, careful prompt design is critical. In this work, we examine several strategies—including prefixing, chain-of-thought (CoT) and dynamic prompting—to evaluate how they enhance the model's internal hidden representations. Specifically, for each attribute α (e.g., brushstrokes, color palette, composition, lighting, abstraction, subject matter), we construct a targeted textual prompt P and a prefix π that emphasizes its visual definition. The VLM then takes the image I along with π and P to generate a response corresponding to the chosen attribute. The set of prefixes employed is listed in Table 1. In this work we consider six attributes—Brushstrokes, Color Palette, Subject Matter, Composition, Lighting, and Abstraction—which together encode the core factors that define an image's stylistic identity.

The selected attributes comprise of content based attributes such as subject matter and composition as well as style based attributes such as lighting and abstraction. We select these attributes primarily based on a) information provided in previous work (Somepalli et al., 2024; Kumar et al., 2025) and b) their complimentary nature encompassing both style and content.

3.2 Attribute Extraction

Visual style is inherently multifaceted, encompassing elements such as color, texture, composition, lighting, abstraction, and subject matter, yet traditional monolithic embeddings flatten these components into a single vector representation (Bai et al., 2021). This semantic compression inhibits interpretability, diminishes stylistic nuance, and often prioritizes content features over stylistic ones, thereby restricting precise control and fine-grained style-aware retrieval. To address these limitations, we propose disentangling style into attribute-specific embeddings.

We begin by conditioning a multimodal vision language model (MLLM) using attribute-specific prompts, yielding an embedding $e_L^{\alpha} \in \mathbb{R}^D$ at layer

L that isolates the stylistic signature of attribute α. Formally, given an image I and an attribute α:

$$\{e^{\alpha}_{L_1},\ e^{\alpha}_{L_2},\ \ldots,\ e^{\alpha}_{L_N}\} = \mathrm{MLLM}\big(I, \pi(\alpha),\ P(\alpha)\big),$$

where $\pi(\alpha)$ and $P(\alpha)$ denote the prefix and prompt associated with attribute α, providing the appropriate context to ensure that the MLLM extracts the relevant stylistic information. The resulting structured representation

$$E^{\alpha}(I) = \{e^{\alpha}_{L} \mid \alpha \in A,\ L \in \{1, 2, \ldots, N\}\},$$

captures each stylistic dimension independently across layers. Additionally, we also incorporate the vision encoder's representation $E^{v}(I) = \mathcal{E}_v(I)$, where $\mathcal{E}_v$ denotes the vision encoder (typically a ViT) that serves as the "eye" of the MLLM.

3.2.1 Dynamic Prompting and Prefixing

Multimodal Large Language Models (MLLMs) can be steered through prompts and system prefixes; however, designing task-relevant prompts manually is often time-consuming and suboptimal. To address this, we propose an automated framework that jointly generates a prompt P and a prefix π conditioned on the input image and task. We illustrate this in Figure 5.

Student–Critic Prompt Generation. Let I denote an input image and T a downstream task. We employ a student–critic interaction to generate a task-adaptive prompt. Let the student MLLM be denoted by S_θ and the critic MLLM by C_ϕ.

At interaction round $t \in \{1, \ldots, k\}$, the student produces a response $y_t = S_\theta(I, \pi, P, h_{t-1})$, where $h_{t-1} = \{y_1, \ldots, y_{t-1}\}$ represents the accumulated interaction history, P is the task specific prompt and π is the system prompt. The critic evaluates the student's response conditioned on the image and conversation history: $c_t = C_\phi(I, y_t, h_{t-1})$, and provides corrective feedback, which is appended to the history for subsequent rounds. After k rounds, we retain the final student response y_k and extract the corresponding hidden representations of downstream retrieval. Specifically, we employ Qwen2.5-VL-7B (Bai et al., 2025) as both the student and the critic for k=3 rounds of interaction. System and the task prompts are described in Figure 5.

Dynamic Prefixing. In addition to prompt content, MLLM outputs are highly sensitive to *instructional prefixes* that condition the model's reasoning behavior. A prefix differs from a prompt in that it does not specify the task itself, but instead provides a meta-level instruction that biases how the model attends to visual evidence and structures its response. To exploit this, we introduce a prefix agent A_ψ that generates a task-aware *user-level prefix*, which is prepended to the main query as a user prompt and emphasizes visual cues. This prefix functions as a soft conditioning prior that consistently steers the model's visual reasoning along with the prompt.

Given the image I and task description T, the prefix agent produces: $\pi = A_\psi(I, T)$. The final inference is performed by conditioning the student model on the generated prefix and prompt: $y = S_\theta(I, P, \pi)$ (please refer supplementary for exact prompt format). The processing and retrieval methodology is the same as that used for automatic prompt generation.

Table 2: Retrieval performance (%) for different components across attributes.

Prompt	Recall@1	Recall@10	mAP@10
Abstraction	56.63	91.45	60.52
Brushstrokes	56.94	91.70	60.54
Color Palette	57.02	91.58	60.48
Composition	56.63	91.66	60.51
Lighting	56.77	91.70	60.54
Subject Matter	56.69	91.54	60.52
Mean	56.78	91.60	60.52
Voting	**58.19**	**92.60**	**64.52**

3.3 Retrieval

For retrieval, we extract layer-wise embeddings from the vision–language backbone and compute their mean across the token dimension, following (Sun et al., 2025). For each layer, we perform retrieval using cosine similarity and nearest-neighbor search, producing a ranked list of retrieved images that reflects the layer's representational bias. To form the final answer, we select the image that appears most frequently across the retrieved lists of all layers, effectively performing a majority-vote aggregation. This approach leverages cross-layer consensus without introducing additional learnable parameters and produces more discriminative retrieval results than relying on any single layer.

4 Results

We evaluate our approach and compare against several state-of-the-art visual encoders and style-specific models. We conduct our experiments on subsets of the WikiArt (Saleh and Elgammal, 2015) and DomainNet (Peng et al., 2019) datasets. For

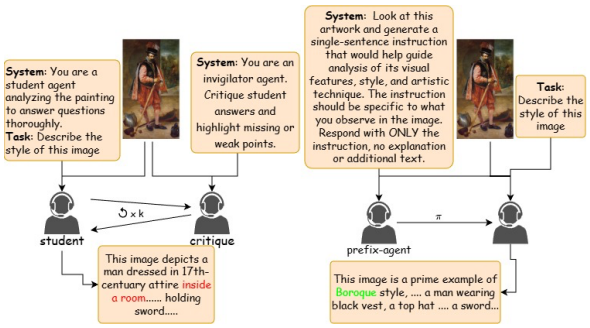

Figure 5: Dynamic Prompting (Left) vs. Dynamic Prefixing (Right). Prefixing generates an automated user-level instruction that conditions model reasoning prior to the prompt, eliminating manual prefix and prompt design.

WikiArt, we randomly sample 200 database images and 100 query images from each of the 27 categories, resulting in a chance probability of 1/27. Similarly, for DomainNet, we sample 500 database images and 50 query images per class. We use the class labels as proxy for the attribute label. As reported in Table 3, traditional representations such as VGGGram (Gatys et al., 2015) and ResNet-50 (He et al., 2016) achieve moderate retrieval performance, with mAP@10 ranging from 52.06% to 54.38% on WikiArt and 65.73% to 72.23% on DomainNet. Self-supervised vision transformers, including DINO ViT-B/8 and CLIP ViT-B/16, improve performance substantially, particularly on Recall@1 and Recall@10 metrics, demonstrating the benefits of learned feature representations for style-aware retrieval. SigLIP models (Zhai et al., 2023), despite their scale, underperform compared to both traditional and self-supervised baselines.

The CSD baseline (Somepalli et al., 2024), which is specifically trained for style retrieval, achieves 62.00% mAP@10 and 57.93% Recall@1 on WikiArt, and 75.60% mAP@10 and 72.80% Recall@1 on DomainNet. Our method, Qwen2.5-VL-7B, slightly improves these numbers (64.52% mAP@10 and 58.19% Recall@1 on WikiArt; 97.23% mAP@10 and 99.86% Recall@1 on DomainNet), but unlike the baselines, it is completely training-free and enables attribute-level matching (modular representations) rather than overall style matching (monolithic representations). This attribute-aware capability allows interpretable and fine-grained retrieval, demonstrating that our approach provides both competitive performance and additional functionality without any additional training. We also perform per-attribute retrieval and present the results in Table 2 and Figure 4.

Results on Other Models. We conduct the same analysis on two additional vision–language models: LLaVA-1.6-Mistral-7B (Liu et al., 2024) and InternVL3-7B (Chen et al., 2024). Averaged across all attributes, InternVL3 achieves a mean accuracy of 52.69%, while LLaVA-1.6 attains 51.63%. We attribute this lower performance primarily to weaker vision encoders: the InternVL visual encoder achieves an accuracy of 50.86% and the LLaVA-1.6 visual encoder 50.42%, compared to 56.13% for the Qwen2.5 ViT on WikiArt. Despite the overall performance gap, our core findings and trends remain consistent across these models.

Retrieval tasks in BLINK benchmark. Table 4 presents the accuracy of three tasks from the BLINK benchmark (Fu et al., 2024), namely Art Style, Visual Similarity, and Jigsaw. Text accuracy is computed based on the answers generated by the model. To compute nearest-neighborhood (NN) accuracy, we extract the hidden states of the model for the query image and the option images along with the prompt explaining the task, then identify the nearest option for each layer of the VLM and use hard-voting scheme across all layers to determine the final answer. Across all tasks, NN accuracy consistently outperforms text accuracy by a considerable margin, highlighting the utility of MLLM hidden states for retrieval and matching tasks.

5 Human preference study

We conducted a human preference study to evaluate attribute-based image retrieval. For each of the six attributes, participants answered five questions, resulting in a total of thirty questions. Each question presented a query image along with three retrieved images: one from our method and one each from CLIP and CSD. The study was conducted in a blind setup, meaning participants were unaware of which method produced each image. 22 participants took part in the study ($22 \times 30 \times 3 = 1980$ votes). Our method was preferred 54.95% of the time, compared to 22.61% for CLIP and 22.45% for CSD. These results demonstrate that our attribute-specific retrieval approach aligns more closely with human perception and underscore the limitations of monolithic representations, which tend to retrieve a single object irrespective of contextual or attribute-specific nuances.

Model	WikiArt			DomainNet		
	R@1	R@10	mAP@10	R@1	R@10	mAP@10
VGGGram (Gatys et al., 2015)	34.06	77.07	53.19	62.80	94.00	70.75
DINO ViT-B/8 (Caron et al., 2021)	46.56	85.33	60.44	68.80	93.20	74.86
DINO ViT-B/16 (Caron et al., 2021)	44.94	84.45	59.87	61.60	93.60	72.86
CLIP ViT-B/16 (Radford et al., 2021)	53.94	89.87	64.11	64.00	96.80	71.01
ResNet-50 (He et al., 2016)	35.68	79.87	54.38	66.80	94.40	72.23
SigLIP ViT-SO400M/14 (Zhai et al., 2023)	31.89	72.78	53.09	58.40	91.20	66.27
SigLIP ViT-L/16 (Zhai et al., 2023)	32.01	76.12	52.06	50.40	90.40	65.73
SigLIP ViT-B/16 (Zhai et al., 2023)	30.55	74.86	51.36	51.20	93.20	64.30
CSD (Somepalli et al., 2024)	57.93	92.45	62.00	72.80	97.60	75.60
Qwen2.5-VL-7B (Ours)	58.19	92.60	64.52	99.86	100.00	97.23

Table 3: Image retrieval performance (%) on WikiArt and DomainNet datasets. Our approach consistently outperforms generic vision encoders as well as models specifically trained for style retrieval, despite requiring no additional training.

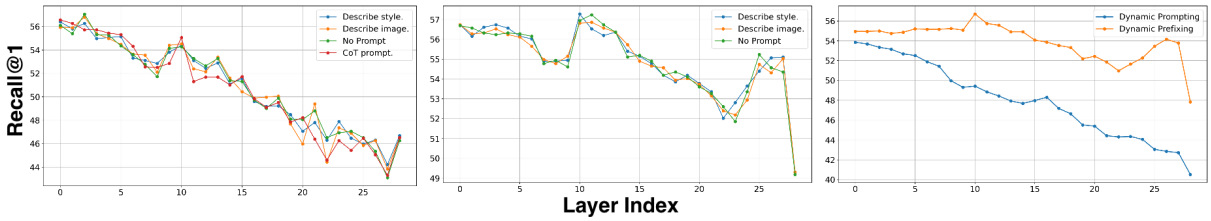

Figure 6: (Left) Recall@1 (accuracy) across layers showing information degradation for different prompts, including a chain-of-thought prompt. (Middle) Middle layers capture richer hidden representations when conditioned on an appropriate prefix. (Right) Performance comparison of automated versus manual prompts and prefixes, showing that automated prompts, just like manual prompts, also experience progressive information loss in deeper layers.

Model	Art Style		Visual Similarity		Jigsaw	
	Text	NN	Text	NN	Text	NN
Qwen2.5-VL-3B	58.1	70.9	57.0	83.0	50.0	70.7
Qwen2.5-VL-7B	68.4	78.6	85.2	85.9	63.3	75.3
Qwen2.5-VL-32B	70.1	71.8	79.3	83.0	68.0	72.0
InternVL3-2B	47.0	70.1	54.7	73.5	67.5	69.2
InternVL3-8B	51.9	82.2	52.6	86.7	64.4	85.9
InternVL3-14B	46.7	80.0	47.3	82.0	57.3	81.3

Table 4: Text-based vs. nearest-neighbor (NN) accuracy on three retrieval tasks from the BLINK benchmark. NN accuracy is computed using MLLM hidden-state embeddings with layer-wise hard voting, consistently outperforming text-based answers across all tasks.

6 Experiments and Analysis

Attributes encapsulate complimentary information. To analyze whether the attributes capture complementary information, we preprocessed all textual responses generated by the MLLM by removing stop words, punctuation, and numbers, and computed Jaccard similarity and TF-IDF cosine similarity across the six attributes for each query sample, reporting the mean values across all samples. This is illustrated in Figure 7. Unlike prior works (Kawarada et al., 2025) that force the model to output a single token per attribute, our method aggregates embeddings across all tokens in the LLM response; directly comparing these pooled embeddings would be dominated by common tokens and, therefore, not reliably indicate attribute-specific information. Instead, our token-level analysis shows that the mean Jaccard similarity between attributes is extremely low (off-diagonal values mostly between 0.03–0.09), indicating minimal overlap in token usage, while the mean TF-IDF cosine similarity is also low (off-diagonal values below 0.14), demonstrating that the relative importance and distribution of words differ across attributes. We observe slightly higher similarity values for subject matter and composition pair, as both are content-specific attributes that describe the objects in the image and therefore naturally share more tokens. These results indicate that each attribute provides a largely distinct textual signal, suggesting that embeddings derived from attribute-specific responses encode unique aspects of an image, supporting the use of multi-attribute embeddings for retrieval and other downstream tasks.

Information degradation in shallow layers. Prior work has examined VLM embeddings for visual representation learning (Sun et al., 2025; Zhuang et al., 2024), but the focus has largely remained on the final-layer embeddings. However, the LLM's

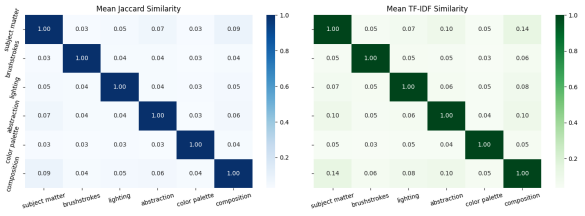

Figure 7: Average Jaccard and TF–IDF similarity between attribute descriptions across all samples, showing low textual overlap and complementary attribute representations.

ability to retain and utilize visual information throughout its layers is known to be a key bottleneck in VLM performance (Fu et al., 2025; Huang et al., 2025). To better understand and validate limitation, we analyze all hidden layers of the LLM to identify the optimal layer for visual feature extraction. We begin by studying knowledge degradation across layers by evaluating retrieval performance at each layer using four prompting strategies: (1) "Describe style in this image," (2) "Describe this image," (3) a chain-of-thought prompt, and (4) no prompt. Figure 6 summarizes the accuracy trends of hidden states across layers. Our results show a clear decay of utilization of visual information in deeper layers (10% mAP@1 decrease), confirming that VLMs tend to suppress image features as language reasoning dominates. Additionally, the sensitivity—measured as the influence of prompt on the hidden states—is minimal when averaged across tokens, suggesting that deeper-layer activations are increasingly steered by the text rather than visual embeddings. Moreover, we also test the retrieval performance on arbitrary prompts including unrelated text such as 'A man stands next to a car' and a no-prompt condition that allows the VLM to produce an unconstrained response. Across all cases, we observe that the influence of the prompt on the final-layer hidden states is minimal, with only marginal variation in mAP and recall as the model essentially ignores the prompt and instead provides the description of the image.

Prefixing enriches intermediate embeddings. Prefixing has been shown to improve VLM performance for correspondence tasks (Liu et al., 2025) by influencing the visual tokens in the KV cache of the transformer head. However, its effect on hidden states for retrieval tasks remains largely unexplored. We begin by introducing a relevant prefix and observe substantial improvements in the performance of intermediate layers, indicating that these layers are highly sensitive to contextual cues. Figure 6 (Bottom) illustrates the per-layer performance when a prefix is used to disentangle the image. We observe that automated prefixing outperforms dynamic prompting by 5.9% averaged across all layers. Moreover, it also outperforms manual prompting (See Table 5).

Table 5: Retrieval performance (%) using different Qwen2.5-based configurations.

Method	Recall@1	Recall@10	mAP@10
Qwen2.5 ViT	56.13	90.86	60.36
Qwen2.5 + P (Mean)	56.52	90.96	60.36
Qwen2.5 + (π,P (Mean),α)	56.82	91.52	60.38
Qwen2.5 + (π,P,Voting,α)	**58.14**	**92.86**	**62.22**

Ablation on model components. Table 5 demonstrates the incremental effect of different components in our retrieval framework. Starting from the Qwen2.5 ViT baseline, introducing a generic prompt ("Describe style in this image") with mean aggregation across tokens (Qwen2.5 + P) yields a modest improvement in Recall@1 (56.13 $\rightarrow$ 56.52), indicating that prompting alone provides limited but consistent gains. Adding the proposed prefixing technique and attribute-specific extraction (π, P, α) further improves performance, suggesting better alignment of attribute-specific representations. Finally, voting-based fusion across all transformer layers boosts performance to a Recall@1 of 58.14, outperforming all baselines.

7 Conclusion

We introduce a training-free framework for attribute-based retrieval, which represents each artwork through six interpretable, attribute-specific embeddings extracted via Multimodal Large Language Models. Our analysis reveals that 1) intermediate MLLM layers preserve richer visual information than the visual encoder and final-layer outputs; 2) our proposed dynamic prefixing significantly outperforms manual prompt design; and 3) Layer-wise hard-voting fusion aggregates evidence across transformer layers, outperforming models specifically trained for style retrieval on WikiArt and DomainNet datasets. Beyond artistic style retrieval, our framework enables zero-shot attribute specification for content moderation, dataset curation, and generative model evaluation. By demonstrating that MLLMs can perform fine-grained visual retrieval without task-specific training, we hope to inspire further exploration of hidden representations for open-set visual understanding.

8 Limitations

While our method is effective and flexible, it has a few limitations. Although training-free, it relies on large pretrained MLLMs and multi-layer hidden-state extraction, which introduces a higher inference cost compared to lightweight pretrained vision encoders. Further, attribute separation is achieved through prompting rather than explicit architectural constraints, so some overlap can occur between closely related attributes, particularly the ones tied to semantic content.

References

Adobe. 2023. Firefly. https://www.adobe.com/sensei/generative-ai/firefly.html.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zechen Bai, Yuta Nakashima, and Noa Garcia. 2021. Explain me the painting: Multi-topic knowledgeable art description generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *https://cdn.openai. com/papers/dall-e-3.pdf*, 2(3):8.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

E. Fernie. 1995. *Art History and Its Methods*. Phaidon Press.

Stephanie Fu, tyler bonnen, Devin Guillory, and Trevor Darrell. 2025. Hidden in plain sight: VLMs overlook their visual representations. In *Second Conference on Language Modeling (COLM)*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *18th European Conference on Computer Vision (ECCV)*.

Siddhartha Gairola, Rajvi Shah, and PJ Narayanan. 2020. Unsupervised image style embeddings for retrieval and recognition tasks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *Journal of Vision*.

Daniel J Graham, James M Hughes, Helmut Leder, and Daniel N Rockmore. 2012. Statistics, vision, and the analysis of artistic style. *Wiley Interdisciplinary Reviews: Computational Statistics*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kung-Hsiang Huang, Can Qin, Haoyi Qiu, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2025. Why vision language models struggle with visual arithmetic? towards enhanced chart and geometry understanding. In *Findings of the Association for Computational Linguistics*.

James M Hughes, Daniel J Graham, and Daniel N Rockmore. 2010. Stylometrics of artwork: uses and limitations. In *Proceedings of Computer Vision and Image Analysis of Art*.

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*.

Masayuki Kawarada, Kosuke Yamada, Antonio Tejero-de Pablos, and Naoto Inoue. 2025. Training-free conditional image embedding framework leveraging large vision language models. *arXiv preprint arXiv:2512.21860*.

Anand Kumar, Jiteng Mu, and Nuno Vasconcelos. 2025. Introstyle: Training-free introspective style attribution using diffusion features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Sara Lawrence-Lightfoot and Jessica Hoffmann Davis. 2002. *The art and science of portraiture*. John Wiley & Sons.

Adrian Lecoutre, Benjamin Negrevergne, and Florian Yger. 2017. Recognizing art style automatically in painting with deep learning. In *Asian conference on machine learning*. PMLR.

Jia Li, Lei Yao, Ella Hendriks, and James Z Wang. 2011. Rhythmic brushstrokes distinguish van gogh from his contemporaries: findings via automated brushstroke extraction. *IEEE transactions on pattern analysis and machine intelligence*.

Benlin Liu, Amita Kamath, Madeleine Grunde-McLaughlin, Winson Han, and Ranjay Krishna. 2025. Visual representations inside the language model. *Second Conference on Language Modeling (COLM)*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Jieyi Long. 2023. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*.

Shin Matsuo and Keiji Yanai. 2016. Cnn-based style vector for style image retrieval. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*.

Orfeas Menis-Mastromichalakis, Natasa Sofou, and Giorgos Stamou. 2020. Deep ensemble art style recognition. In *2020 International Joint Conference on Neural Networks (IJCNN)*.

Midjourney. n.d. Midjourney. https://www.midjourney.com/home.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415.

Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. 2023. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *Proceedings of The Twelfth International Conference on Learning Representations (ICLR)*.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. 2022. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Dan Ruta, Saeid Motiian, Baldo Faieta, Zhe Lin, Hailin Jin, Alex Filipkowski, Andrew Gilbert, and John Collomosse. 2021. Aladin: all layer adaptive instance normalization for fine-grained style similarity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11926–11935.

Robert Sablatnig, Paul Kammerer, and Ernestine Zolda. 1998. Hierarchical classification of paintings using face-and brush stroke models. In *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, volume 1, pages 172–174. IEEE.

Babak Saleh and Ahmed Elgammal. 2015. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *International Journal for Digital Art History*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. 2024. Investigating style similarity in diffusion models. In *18th European Conference on Computer Vision*.

Bhargav Srinivasa Desikan, Hajime Shimao, and Helena Miton. 2022. Wikiartvectors: style and color representations of artworks for cultural analysis via information theoretic measures. *Entropy*, 24(9):1175.

Li Sun, Chaitanya Ahuja, Peng Chen, Matt D'Zmura, Kayhan Batmanghelich, and Philip Bontrager. 2025. Multi-modal large language models are effective vision learners. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8617–8626.

Joshua Tenenbaum and William Freeman. 1996. Separating style and content. In *Advances in Neural Information Processing Systems*, volume 9.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NeurIPS '22.

Daan Wynen, Cordelia Schmid, and Julien Mairal. 2018. Unsupervised learning of artistic styles with archetypal style analysis. *Advances in Neural Information Processing Systems*, 31.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Yifan Zhang, Yang Yuan, and Andrew Chi-Chih Yao. 2023. Meta prompting for AI systems. *Proceedings of International Conference on Learning Representations Workshops*.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2024. PromptReps: Prompting large language models to generate dense and sparse representations for zero-shot document retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

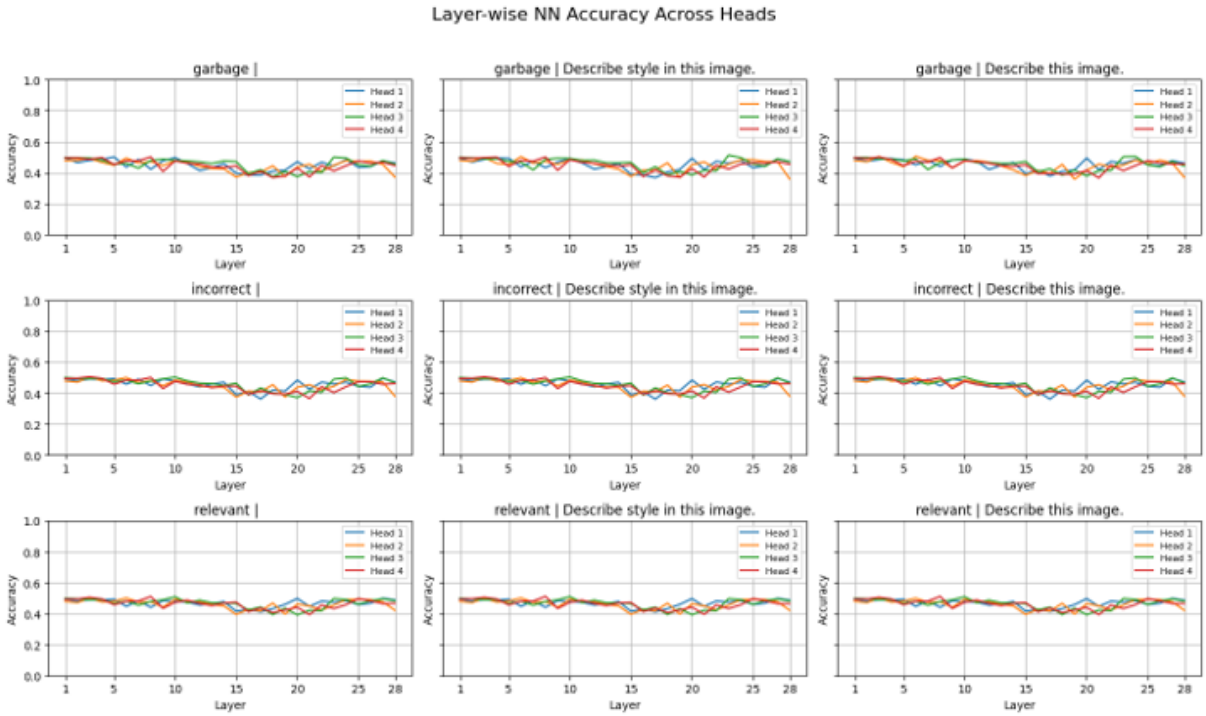

Figure 8: Layer-wise performance of image value tokens in the KV cache for retrieval tasks across 4 attention heads and 3 prompts. Accuracy remains largely unchanged across heads and layers, even when using unrelated or corrupted prefixes (e.g., "A man driving a car" or "hsufhe8...6sjeridh"), indicating that visual tokens do not capture additional style information from the images regardless of the prompt.

9 Supplementary

9.1 Image-value tokens as representation:

In (Liu et al., 2025), the authors explore the role of visual tokens in the KV cache as the image value tokens act as a proxy for the input image at each layer of the language model. As illustrated in Figure 8, we analyze these image value tokens across all heads and layers for retrieval tasks and find that their performance remains largely unchanged across layers, regardless of the prompts used. This suggests that the VLM does not leverage additional style information from the images based on the prompt, making visual tokens an ineffective choice for representing style.

9.2 Ablation of Fusion operator.

While our main approach employs hard voting across all layers to highlight the non-parametric utility of MLLM embeddings, we also evaluate trainable alternatives. Specifically, we fit a random forest and a logistic regression on the database embeddings at the top-performing layer (layer 10, Figure 5) and report mean accuracy across all attributes. The random forest achieves 62.17% and logistic regression achieves 64.54%, compared to 58.19% using hard voting. This indicates that intermediate embeddings can be potentially leveraged by lightweight trainable heads to improve retrieval performance.

9.3 Prompt formats

Information Degradation

```
messages = [
    {
    "role": "user",
    "content": [
      {"type": "image", "image": img_path},
       {"type": "text", "text": prompt}
    ]
    }
]
```

where prompt is ‘" "‘, ‘"Describe this image"‘, and ‘"Describe style in this image"‘.

Table 6: Recall and mAP at different retrieval thresholds (1, 10, 100) for various parameters and methods without prefixing.

Parameter / Method	**Recall (%)**			**mAP (%)**		
	1	10	100	1	10	100
CSD	57.93	92.45	99.79	57.93	62.00	43.68
Qwen ViT	56.13	90.86	99.79	56.13	60.36	41.54
"Describe style of this image"	45.85	85.62	99.25	45.85	52.83	34.71
"Describe this image"	46.48	85.79	99.29	46.48	52.54	34.70
Empty prompt	46.52	85.99	99.21	46.52	52.70	34.71
Random prompt	46.27	84.49	99.42	46.27	52.08	34.36
Abstraction	44.35	81.45	99.25	44.35	55.22	37.77
Brushstrokes	44.27	81.70	99.42	44.27	55.21	37.64
Color Palette	43.14	81.24	99.25	43.14	55.09	37.58
Composition	43.31	81.62	99.29	43.31	55.09	37.63
Lighting	44.02	81.37	99.42	44.02	55.64	37.75
Subject Matter	43.60	81.28	99.33	43.60	55.49	37.73

Figure 9: An example of attributes described by MLLM.

Prompt for Prefixing

```
messages = [
    {
    "role": "system",
   "content": "You are an art curator
           analyzing this painting.
            Analyze visual features
           that reveal the painter's
            technique and aesthetic
                 choices."
    },
    {
    "role": "user",
    "content": [
        {"type": "image",
            "image": img_path},
        {"type": "text",
            "text": prompt}
    ]
    }
]
```

where prompt is ‘" "‘, ‘"Describe this image"‘, and ‘"Describe style in this image"‘.

10 Human Survey Analysis

Table 7 presents the survey results comparing our method with CLIP and CSD across six stylistic attributes. Prior methods rely on holistic embeddings that conflate multiple stylistic dimensions and cannot distinguish between attributes. In contrast, our approach decomposes visual style into attribute-specific representations, enabling interpretable comparisons. This design is consistently preferred by humans across all attributes, with particularly strong margins for lighting and subject matter, demonstrating that attribute disentanglement improves both interpretability and perceptual alignment.

Table 7: Per attribute human preference accuracies.

Attribute	Ours	CLIP	CSD
Abstraction	53.4	36.4	10.2
Brushstrokes	63.6	20.9	15.5
Color Palette	54.5	18.2	27.3
Composition	56.8	14.8	28.4
Lighting	70.6	16.5	12.8
Subject Matter	67.3	10.0	22.7

Prompt for Dynamic-Prefixing

Prompt to prefix generator:

```
messages = [
    {
    "role": "user",
    "content": [
        {"type": "image",
            "image": img_path},
        {"type": "text",
        "text": "Look at this artwork
        and generate a single-sentence
            instruction that would
            help guide analysis of
            its {attribute}.
            The instruction
         should be specific to what
           you observe in the image
          regarding {attribute}."}
    ]
    }
]
```

Prompt to embedding generator:

```
messages = [
    {
    "role": "system",
    "content": GENERATED_PREFIX
    },
    {
    "role": "user",
    "content": [
        {"type": "image",
            "image": img_path},
        {"type": "text",
            "text": prompt}
    ]
    }
]
```

where prompt is described in Table 1.

Figure 10: Example of samples used in human study. The image order is randomized during the study.