WHAT DOES A VISUAL FORMAL ANALYSIS OF THE WORLD'S 500 MOST FAMOUS PAINTINGS TELL US ABOUT MULTIMODAL LLMS?

Muzi Tao *
University of Southern California
muzitao@usc.edu

Saining Xie
New York University
saining.xie@nyu.edu

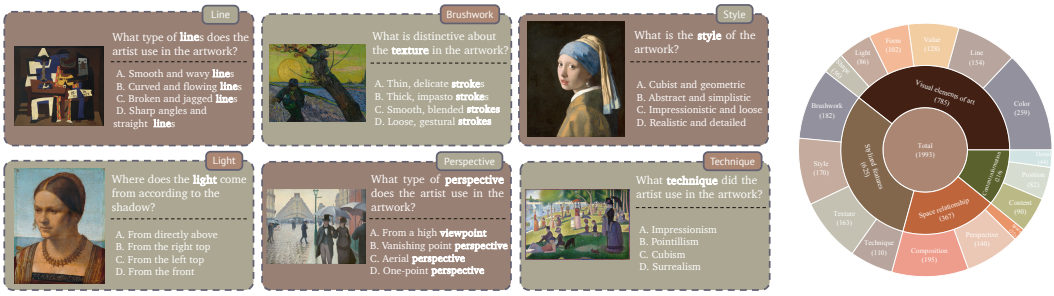


Figure 1: To truly understand and appreciate art, it is essential to use our vision. We focus on the visual *formal analysis* and meticulously explore the arrangement and function of visual elements of art. Through visual formal analysis, we can comprehend the artistic and emotive essence of a piece, independently of its historical background or the artist's motives. The process of formal analysis in art transforms visual impressions into articulate verbal descriptions, offering a vision-centric environment for evaluating multimodal systems.

ABSTRACT

This work introduces ArtQA, a new benchmark for multimodal LLMs through the lens of *formal analysis* of paintings. We focus on key elements such as *line, shape, space, color, form, value, and texture*—collectively referred to as the elements of art in visual formal analysis. ArtQA contains questions spanning 4 metrics, further divided into 16 fine-grained categories. We leverage the power of LLMs to generate VQA questions based on the formal analysis of 500 renowned paintings. These questions undergo a rigorous filtering process by both model annotation and human experts, ensuring ArtQA's quality and reliability.

1 INTRODUCTION

Recently, there has been rapid development in large language models (LLMs) (OpenAI, 2023c; Touvron et al., 2023), which have paved the way for the emergence of multimodal large language models (MLLMs). By processing additional image input and performing visual question answering (OpenAI, 2023a; Liu et al., 2023b), MLLMs showcase impressive proficiency in handling non-trivial visual tasks. Yet, a pertinent question remains: does the vision component play a crucial role, or is it simply offering visual context for the LLMs to do the heavy-lifting? We contend that to fully evaluate the multimodal capabilities of these models, we need to increase the bar—it's essential to pinpoint situations where precise perception isn't just advantageous, but absolutely necessary. However, a significant challenge arises from the fact that detailed visual processing of objects is often nuanced and implicit in nature. This subtlety presents difficulties in distinctly identifying and expressing these processes in clear, straightforward language, and turning them into evaluation questions. We recognize *art*, particularly the formal analysis of art, as an effective tool for communicating these intricate visual characteristics. The process of formal analysis in art transforms visual impressions into articulate verbal descriptions, aligning perfectly with our goal to associate specific visual element features with their corresponding descriptive text.

To this end, we have developed a new benchmark to evaluate MLLMs, named **ArtQA**, which utilizes *formal analysis* to focus on critical elements such as line, shape, space, color, form, value, and texture, referred to as the elements of art. The benchmark comprises 1993 multi-choice questions from 500

*Work done during an internship at NYU.

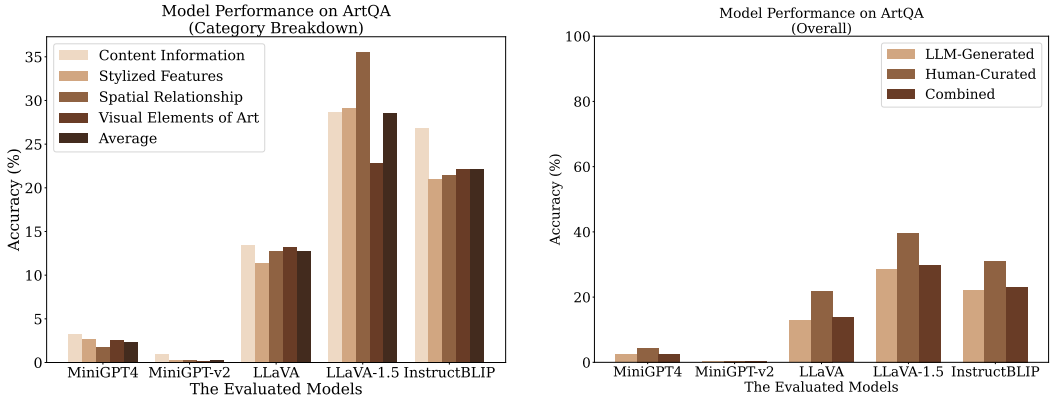


Figure 2: Model Performances on ArtQA. MLLMs still struggle with visual formal analysis questions.

famous artworks. To ensure the quality and reliability of ArtQA, we have implemented a dual-layer filtering processing with both model selection and human-expert annotation. In our evaluation of State-of-the-Art models on ArtQA, we found, unsurprisingly, that current MLLMs still struggle with these intricate visual details.

2 Constructing ArtQA

Question Generation We use the top 500 famous paintings. Most of the images are from WikiArt (Saleh & Elgammal, 2015), wikipedia (wikipedia, 2023), and other artist's websites. The whole pipeline of ArtQA construction is presented in Figure 6.

We utilize GPT-4 as a tool for gathering professional opinions about artworks. Generally, the prompting process consists of two parts. The first part focuses on generating formal analysis. In the second phase of our process, the objective is to create questions based on previously acquired formal analyses. Below is an example of our prompt for the second part of our process.

Data Filtering To make sure we generate high-quality and reliable VQA questions on formal analysis, we introduce a two-step filtering process. The first stage involves model annotation to eliminate questions that exhibit textual bias. We applied standard models to questions generated by GPT-4 for VQA and discarded those where more than two models provided the correct annotation. This initial filtering step effectively reduces the question pool by about half, leaving 2,489 questions for further analysis, down from the original 5,000. Then, we further employ human experts to only keep the good data. See the example of human interface UI in Figure 5. With the interface, there have been 2,489 questions before annotation. And after filtering and correcting, there are 1,993 questions.

3 ArtQA Benchmark

For the 1993 questions in ArtQA, we use GPT-4 to re-categorize them into different categories. The categories and the distribution of the questions appear on the right of Figure 1).

We evaluate the performance of multimodal language models on ArtQA, both on GPT-4 generated questions and human-designed questions. We choose a circular strategy as our evaluation method. Here indicates the performance of selected representative models in the following Figure 2. From the data shown in the figure, we see most of the models are capable of solving these questions in the art domain, yet have a large space to be improved. Moreover, the same rank and almost identical changing curve of the model performance on GPT-4 generated questions and human-designed questions indicate that the part of GPT-4 generated questions serve the same function as human-designed questions. Also, we randomly and uniformly selected 100 questions from ArtQA to evaluate GPT-4V (OpenAI, 2023a). On these questions, GPT-4V achieves an accuracy rate of 57%, indicating there is still considerable room for improvement.

The detailed performance of the models is shown in Table 2. From the table, a common trend is observed: all models generally underperform in the metric concerning the visual elements of art. On most of the categories of Visual Elements of Art metric, models show worse performance than the overall evaluation. Given that these elements are fundamental to vision and represent a key aspect of the detailed visual processing of objects, this shortcoming highlights the existing gaps between detailed vision processing and high-level semantic understanding.

URM STATEMENT

Author Muzi Tao meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 10

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a. 10

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b. 1, 10

OpenAI. Gpt-4v(ision) system card, 2023a. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf. 1, 2

OpenAI. Gpt-4 technical report, 2023b. 4

R OpenAI. Gpt-4 technical report. *arXiv*, pp. 2303–08774, 2023c. 1

Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015. 2

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

wikipedia. wikipedia, 2023. URL https://www.wikipedia.org/. 2

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 10

A DETAILS ABOUT EVALUATION METRICS

In Section 3, we provide the evaluation metrics and the fine-grained categories. Table 1 presents the detailed definition of the categories. Examples of each specific category are shown in Figure 3.

B QUESTION GENERATION DETAILS

In Section 2, we display the construction pipeline process of ArtQA. In the part of question generation, we present how to use prompt with GPT-4OpenAI (2023b) to generate high-quality questions. Here are the complete prompts for the two stages.

Stage 1:

> Provide a formal analysis on the artwork ` ` `{"artist": <*artist*>, "title": <*title*>} ` ` ` using about 500 words.
> In the formal analysis you provide, please use the words "the artist" to replace all the actual name of the artist and use the words "the artwork" to replace all the actual name of the artwork.

Stage 2:

> Design 10 questions each with 4 options to choose from according to the formal analysis below delimited by triple backticks.
> The questions are supposed to focus on how specific representation of a visual element or perspective of formal analysis is in the artwork ` ` `{"author": <*artist*>, "title": <*title*>} ` ` `.
> To generate the questions do the following:
> step 1 - First, choose a perspective from formal analysis to do with style, technique, line, shape, form, texture, color and other elements of art, then find a description of that perspective but do not choose descriptions to do with artwork scale, background knowledge and symbolism.
> step 2 - Second, transfer the description into a question format with a concise correct answer. The answer part should be concise descriptive words. The questions are supposed to focus on how specific the representation of a visual element or perspective of formal analysis is.
> step 3 - Third, randomly choose one option label from (A, B, C, D) to fill in with the correct answer. Then, provide other option labels with answers which are wrong but also possible description words when referring to this question of formal analysis. The lengths of the four options should be similar.
> step 4 - Turn to step 1 until the index of the question reaches 10.
>
> Attention: If the name of an artist or artwork appears in the question or options, please replace them with the words "the artist" and "the artwork".
>
> ` ` `<*analysis*> ` ` `
>
> Provide them in a JSON object with the following JSON format below.
> Format:
> { "Title": <*title*>, "Artist": <*artist*>,
> "Questions": [{
> "index": 1,
> "step1": { "Perspective": <perspective>,
> "Description": <description> },
> "step2": { "Question": <question>,
> "Answer": <correct answer> },
> "step3":{ "Question": <question>,
> "Option A": <option A>,
> "Option B": <option B>,

"Option C": <option C>,
"Option D": <option D>,
"Option": <correct option>,
"Answer": <correct answer> } }, ...] }

The *italicized words* represents the changeable parts of prompts. Specifically, they are the input variables of the artist and title which correspond with the identification of each artwork in both stages, and the input text of analysis in stage 1 which is generated from stage 2.

For the GPT model used in the question generation, we called the API of the GPT-4 model of the 0613 version. Using the same prompts on the updated version might obtain more desirable question results with its training data consisting of a broader knowledge scope.

C EXAMPLES OF ARTISTIC ANALYSIS ON ARTQA QUESTIONS

To cast a deeper insight into the understanding of the relationship between formal analysis and ArtQA questions, Figure 7 presents some examples of artistic analysis of the questions from ArtQA. These questions are related to the professional art domain knowledge and focus on visual information. The column on the right of Figure 7 justifies the correct answer to the questions.

D DATASET VISUALIZATION

We visualize the words of question and the options from the four metrics into word clouds, seen in Figure 4. In the word cloud images, the size of each word indicates its frequency or importance. Frequently used words are typically larger and more centrally located, while less frequent words are smaller and placed around the edges. The word cloud images reveal that the four metrics focus on different yet essential elements in the detailed visual processing of objects.

E MODEL PERFORMANCE ON EVALUATION METRICS

The left figure in Figure 2 presents the model performance on the four different metrics. Each metric provides an evaluation of different aspects of artistic formal analysis. The figure shows that the accuracy on each metric is balanced, further showcasing the balancing design in ArtQA.

F MORE EVALUATION RESULTS ON *SOTA* MODELS

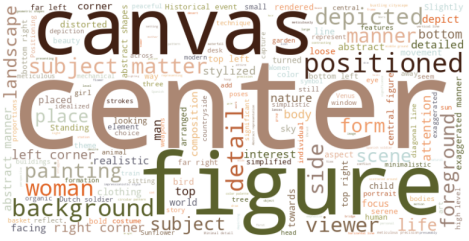
The model performance across each category is also detailed in Table 2. This table illustrates that each model exhibits unique strengths and weaknesses in different categories. However, a common trend is observed: all models generally underperform in the metric concerning the visual elements of art. Given that these elements are fundamental to vision and represent a key aspect of the detailed visual processing of objects, this shortcoming highlights the existing gaps between low-level vision processing and high-level semantic understanding. This insight points to areas for future improvement and research, emphasizing the need to bridge the gap between these two crucial aspects of visual interpretation.

G DATASET LIMITATIONS

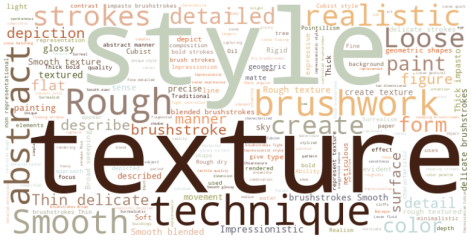
The dataset we've constructed centers around 500 of the most renowned artworks globally. However, this specific focus also presents a limitation regarding the diversity of the artworks included. The selection of these artworks was sourced from a website, which inherently carries some level of bias. Notably, the timeframe of the artworks in the dataset is quite narrow, predominantly featuring pieces created between the years 1400 and 2000. This means that artworks from periods earlier than 1400 are not represented in our collection. Similarly, the dataset doesn't encompass artworks produced in recent years, indicating a gap in contemporary artistic expressions. This exclusion results in a notable limitation, particularly in the representation of modern art styles. Recognizing these limitations, we

Figure 3: 16 Examples of Questions from the ArtQA Benchmark. Each is annotated with a category in the top right corner, and a source in the bottom right corner. The 16 examples are of different categories with the belonged general metric indicated by the outer dashed box. The correct option is colored with light blue.

are actively working towards incorporating a broader range of artworks into the dataset, aiming to include both older and more recent pieces to enhance its diversity and comprehensiveness in the future.



(a) Content Information



(b) Stylized Features

(c) Spatial Relationship

(d) Visual Elements of Art

Figure 4: Word Clouds of the 4 General Metrics.

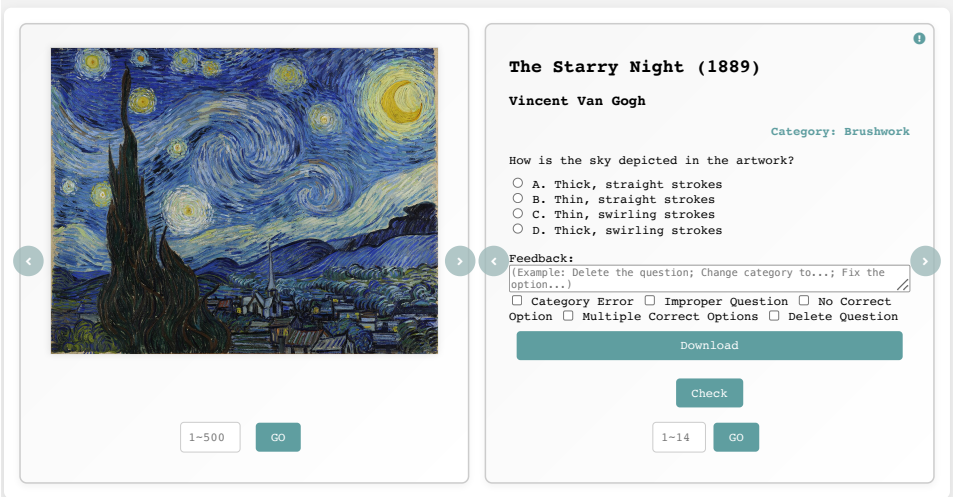


Figure 5: Human Annotation Interface Webpage. The interactive interface collects the answers and the feedback from the users. The left sub-window displays the artwork while the right sub-window displays the question as well as the options. The feedback typed in and the checkbox chosen will be automatically saved.

Category	Definition
Detail	Detail refers to the intricacy and amount of visual information presented. It ranges from broad, simple strokes to finely rendered, complex imagery, significantly influencing the viewer's engagement and interpretation.
Position	Position concerns the placement of elements within a composition. It establishes relationships between objects, contributing to the narrative or thematic structure, and can guide the viewer's eye, creating movement or stability.
Content	Content encompasses the actions, movements, narrative, and thematic elements depicted in the artwork. It narrates what is happening in the scene or what the artwork aims to communicate.
Style	Style is the distinct manner in which an artist expresses their vision, often characterized by specific techniques, color choices, and forms. It varies from being personal to an artist to reflecting a broader art movement.
Texture	Texture refers to the perceived surface quality, beyond just the tactile aspect. It can be visually represented to give an impression of how something might feel, adding depth and realism.
Brushwork	Brushwork denotes the specific methods an artist uses to apply paint with a brush. It can range from smooth, delicate strokes to vigorous, expressive applications, conveying a spectrum of emotions and atmospheres.
Technique	Technique pertains to the various methods and processes artists employ in their creative work. These techniques, varying widely across different media, significantly influence the artwork's final appearance.
Perspective	Perspective is the technique of depicting three-dimensional objects on a two-dimensional surface to create an illusion of depth and space, involving the viewpoint and relative scaling and positioning of objects.
Composition	Composition involves the strategic arrangement of visual elements in an artwork. It includes abstract positioning and scaling of objects to achieve balance, harmony, or intentional discord, often centering around a visual focus.
Space	Space refers to the sense of three-dimensionality and depth created within a two-dimensional medium, distinct from perspective.
Light	Light is used to model forms, create depth, and influence the mood of a piece. It encompasses the direction, intensity, and color of light.
Color	Color involves the study of hues and their relationships, including saturation (intensity) and value (lightness or darkness). It plays a crucial role in creating atmosphere, suggesting mood, and conveying symbolic meanings.
Shape	Shape refers to the two-dimensional aspects of form, defined by contrasting boundaries in color, value, or texture.
Line	Line is utilized for its expressive quality in representing movement, form, and contour. Lines direct the viewer's eye through the composition.
Form	Form represents the three-dimensional aspects of shape, referring to objects with length, width, and depth.
Value	Value includes the lightness or darkness of a color, as well as its tone and temperature. It is essential for creating depth, highlighting focal points, and establishing contrast between light and dark areas.

Table 1: ArtQA Categories and Definitions. The table displays the 16 fine-grained categories and their detailed definitions.

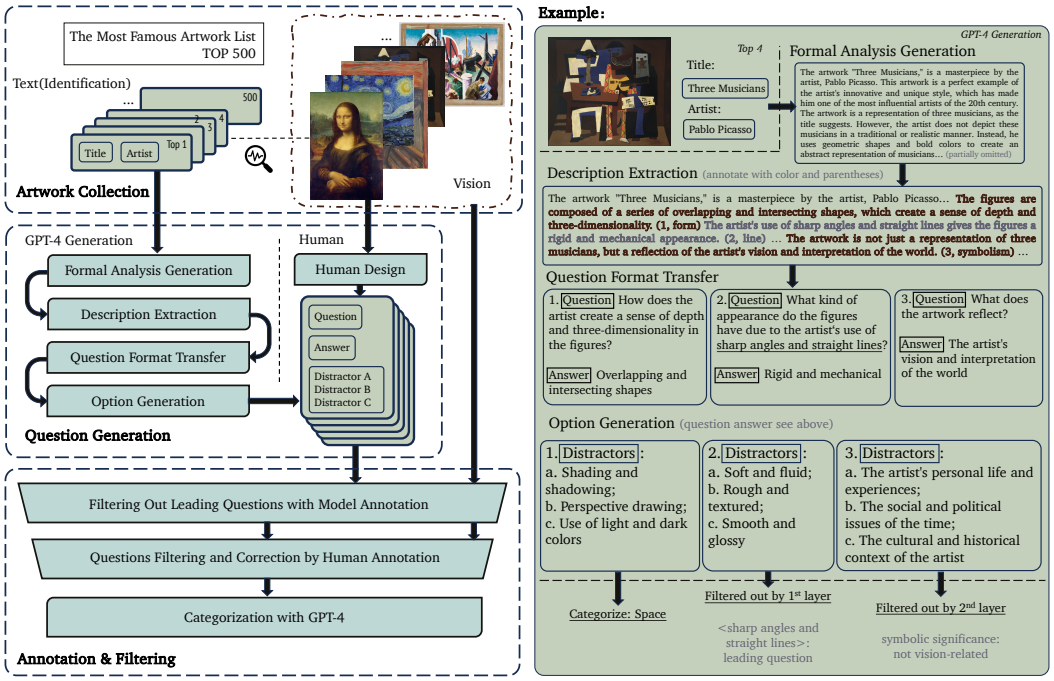


Figure 6: The Construction Process of ArtQA. The left side shows the pipeline of generating questions consisting ArtQA, including artwork collection, question generation, and data filtering. Most of the questions come from the GPT-4 generation by giving simple identification text of the artist and the title. After step-by-step instructed generation, there is two-stage filtering to ensure the high quality of the data. There are also manually designed questions from humans, as part of ArtQA data and the validation of model generated questions. The right side shows an example of the GPT-4 generating pipeline, especially the situations of the filtering process.

Category	Mini-GPT4	Mini-GPT-v2	LLaVA	LLaVA-1.5	Instruct-BLIP
Detail	2.3%	0.0%	22.7%	47.7%	45.5%
Content	2.2%	1.1%	15.6%	23.3%	24.4%
Position	4.9%	1.2%	6.1%	24.4%	19.5%
Space	6.3%	0.0%	15.6%	34.4%	31.3%
Perspective	2.1%	0.7%	14.3%	27.1%	19.3%
Composition	2.6%	0.0%	8.7%	29.7%	20.5%
Style	2.4%	0.6%	14.7%	47.1%	20.6%
Technique	0.9%	0.0%	12.7%	42.7%	25.5%
Texture	2.5%	0.6%	11.7%	34.4%	22.7%
Brushwork	1.1%	0.0%	12.1%	21.4%	18.7%
Value	1.6%	0.0%	21.1%	26.6%	17.2%
Light	2.3%	0.0%	10.5%	24.4%	25.6%
Shape	0.0%	0.0%	10.7%	26.8%	26.8%
Color	1.5%	0.4%	11.2%	25.1%	23.6%
Line	5.2%	0.0%	13.6%	14.3%	24.7%
Form	3.9%	0.0%	11.8%	21.6%	15.7%
Overall	2.4%	0.3%	12.8%	28.6%	22.2%

Table 2: Model Performance on Each Category of ArtQA. Evaluation of five models is presented in the table: Mini-GPT4 Zhu et al. (2023), Mini-GPT-v2 Zhu et al. (2023), LLaVA Liu et al. (2023b), LLaVA-1.5 Liu et al. (2023a), Instruct-BLIP Dai et al. (2023).

Figure 7: Example artistic analysis on questions of ArtQA.