Towards Artwork Explanation in Large-scale Vision Language Models

Kazuki Hayashi[†], **Yusuke Sakai**[†],
Hidetaka Kamigaito[†], **Katsuhiko Hayashi**[‡], **Taro Watanabe**[†]
[†]Nara Institute of Science and Technology [‡]The University of Tokyo
{hayashi.kazuki.hl4, sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp
katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp

Abstract

Large-scale Vision-Language Models (LVLMs) output text from images and instructions, demonstrating advanced capabilities in text generation and comprehension. However, it has not been clarified to what extent LVLMs understand the knowledge necessary for explaining images, the complex relationships between various pieces of knowledge, and how they integrate these understandings into their explanations. To address this issue, we propose a new task: the artwork explanation generation task, along with its evaluation dataset and metric for quantitatively assessing the understanding and utilization of knowledge about artworks. This task is apt for image description based on the premise that LVLMs are expected to have pre-existing knowledge of artworks, which are often subjects of wide recognition and documented information. It consists of two parts: generating explanations from both images and titles of artworks, and generating explanations using only images, thus evaluating the LVLMs' language-based and vision-based knowledge. Alongside, we release a training dataset for LVLMs to learn explanations that incorporate knowledge about artworks. Our findings indicate that LVLMs not only struggle with integrating language and visual information but also exhibit a more pronounced limitation in acquiring knowledge from images alone [1].

1 Introduction

In the field of Vision & Language (V&L), Large Language Models (LLMs) (Touvron et al., 2023; Chiang et al., 2023; Bai et al., 2023a; Jiang et al., 2023) have been combined with visual encoders to create Large Scale Vision Language Models (LVLMs) (Li et al., 2023b; Liu et al., 2024; Bai et al., 2023b; Ye et al., 2023b). These models have achieved success in various V&L benchmarks (Li et al., 2023a; Fu et al., 2023; Liu et al., 2023c; Bai et al., 2023c). Despite these advancements, tasks like Visual Question Answering (VQA) (Zhang et al., 2022b; Yue et al., 2023), Image Captioning(Agrawal et al., 2019; Lin et al., 2014), and querying models about artwork-related information (Garcia et al., 2020; Cetinic, 2021; Bai et al., 2021) have primarily focused on assessing models' abilities to handle isolated pieces of knowledge.

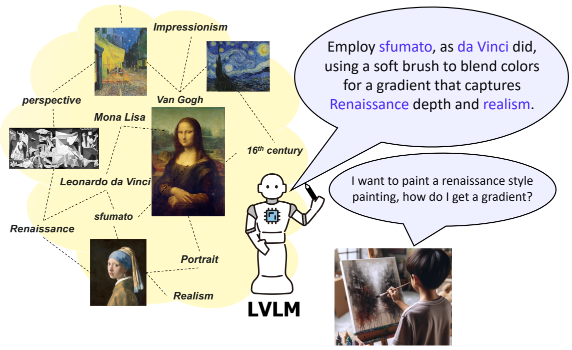

Figure 1: An example of creative assistance using an LVLM, harnessing comprehensive artistic knowledge for guidance.

These tasks, while valuable, do not fully capture the complexity of synthesizing and explaining interconnected knowledge in real-world scenarios (Kawaharazuka et al., 2024), nor the difficulty of generating coherent text to explain this knowledge. Current evaluations often result in superficial image descriptions, lacking extensive background knowledge and interrelationships between subjects.

A pertinent example of this limitation can be observed in the context of creative support for paintings and photographs. As shown in Figure 1, these models must produce explanations that integrate knowledge of the artwork's theme, historical context, associated works, and artistic movement, highlighting a gap in current capabilities. Since this task goes beyond simply recognizing disparate knowledge, it is crucial for LVLMs to deeply understand

[1]The datasets (**ExpArt=Expl**ain **Art**works) are available at https://huggingface.co/datasets/naist-nlp/ExpArt

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 705–729
August 11-16, 2024 ©2024 Association for Computational Linguistics

Type	Template	Instruction	Output
Section	`Explain the {Section} of this artwork, {Title}.`	`Explain the History of this artwork, Mona Lisa.`	Of Leonardo da Vinci's works, the Mona Lisa is the only portrait whose authenticity...
Subsection	`Explain the {Subsection} regarding the {Section} of this artwork, {Title}.`	`Explain the Creation and date regarding the History of this artwork, Mona Lisa.`	The record of an October 1517 visit by Louis d'Aragon states that the Mona Lisa...
Sub subsection	`Explain the {Sub subsection} details within the {Subsection} aspect of the {Section} in this artwork, {Title}.`	`Explain the Creation details within the Creation and date aspect of the History in this artwork, Mona Lisa.`	After the French Revolution, the painting was moved to the Louvre, but spent a brief period in the bedroom of Napoleon (d. 1821) in the....

Table 1: Examples of instructions for the proposed task. The blue part indicates the artwork's title and the red part indicates the names of sections in the original Wikipedia articles that correspond to their explanations.

the interrelationships of artwork knowledge to integrate them into explanations comprehensively.

To address this gap, we propose a new task and evaluation metrics designed to measure LVLMs' capability in generating comprehensive explanations about artworks. Our task requires LVLMs to generate explanations in response to given instructions, based on input images and titles of artworks.

We have constructed a dataset from about 10,000 English Wikipedia articles of artworks for this task and also release a training dataset to facilitate LVLMs in learning to generate explanations involving artistic knowledge. Furthermore, we have evaluated LVLMs currently achieving the highest performance in various V&L benchmarks. The results show that while the LVLMs retain the artistic knowledge inherited from their base LLMs, they do not adequately correlate this knowledge with the provided visual information.

2 LVLMs

LVLMs (Li et al., 2023b; Liu et al., 2024; Bai et al., 2023b; Ye et al., 2023b) integrate a Vision Encoder (Li et al., 2023b) trained through contrastive learning to process visual information with Large Language Models (LLMs) (Li et al., 2023b; Liu et al., 2024; Bai et al., 2023b; Ye et al., 2023b). This integration requires further training to effectively combine vision and language capabilities. As a result, these LVLMs outperform conventional pretrained models, even those with over ten times more parameters (et al, 2022; Driess et al., 2023).

However, it is unclear whether the knowledge from the LLM and the Vision Encoder are appropriately aligned by the additional network layers in LVLMs (Chen et al., 2024a). Generating explanations that involve knowledge about art especially requires careful and systematic alignment and utilization of the information from both the Vision Encoder and the LLM. This challenge motivates us to design a new task for LVLMs.

3 Task and Evaluation Metrics

3.1 Task

Our task demands LVLMs to generate explanations following instructions with images and titles. Examples of the instructions are shown in Table 1. As demonstrated by these examples, each instruction is categorized into three levels, Section, Subsection, and Subsubsection, determined by the corresponding positions in Wikipedia articles (See §3). The proposed task addresses the following two settings with or without titles:

With Title In the context of creative assistance, the title often contains the author's intent for the artwork, and it is desirable to generate explanations considering this intent. In this setting, both the image and its title are inputs, testing whether LVLMs can generate appropriate explanations based on both language and visual information.

Without Title As shown in Figure 1, there are cases where a title does not exist potentially because the artwork is in the process of creation. This setting tests whether LVLMs can generate appropriate explanations using only visual information from images. Additionally, analyzing the performance changes with and without titles allows us to verify the LVLMs' pure vision-based knowledge.

Furthermore, to thoroughly assess the generalization capabilities of LVLMs, we compare two cases: 1) a seen case in which images are observed during finetuning, and 2) an unseen case in which images are not observed during finetuning.

3.2 Evaluation Metrics

Since our task is a kind of natural language generation (NLG), we utilize popular metrics in NLG

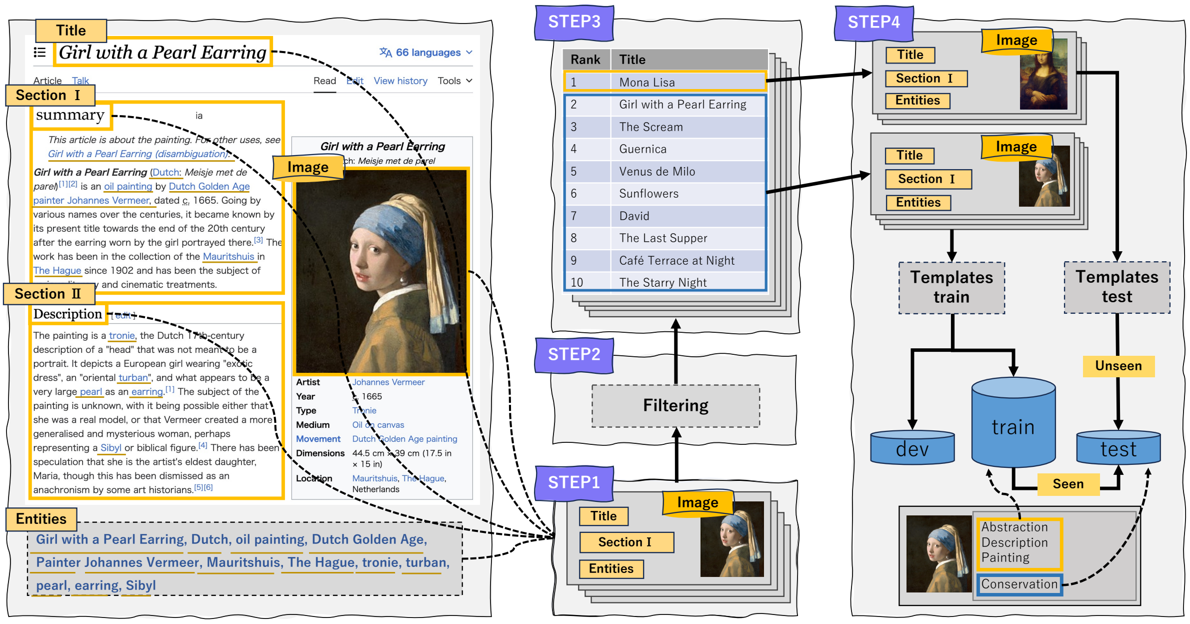

Figure 2: Workflow diagram illustrating the methodology for dataset creation from Wikipedia articles on artworks, involving selection, filtering, data balancing, and instructional templating for LVLM training and evaluation.

	Train	Dev	Test (Seen)	Test (Unseen)
Images	7,704	963	2,407	963
Instruction	18,613	2,677	2,485	2,597

Table 2: Number of Images and Data in the Created Dataset.

for evaluation, i.e., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang* et al., 2020). To further focus on the ability to generate explanations for artworks, we propose the following three evaluation metrics[2]:

Entity Coverage We evaluate how accurately the generated text includes entities (See §4) related to the artwork mentioned in the reference description, using two settings: exact match and partial match (Li et al., 2022a).

Entity F1 We evaluate the frequency of occurrence of entities related to the artwork found in the generated and reference explanations by F1. Inspired by ROUGE, we consider the highest frequency of occurrence of any entities within either the generated explanation or the reference as the upper limit of occurrence frequency to calculate precision and recall.

Entity Cooccurrence This metric assesses not only the coverage of independent entities but also how their interrelations are contextually combined to form the overall explanation. Specifically, it considers pairs of entities that co-occur within a sentence and its preceding and following n sentences, evaluating the coverage rate of these pairs to reveal how well the model understands and integrates the relevance of knowledge. By setting the value of n to exceed the number of sentences in the generated explanation, it becomes possible to account for the co-occurrence of entity pairs throughout the entire text. Furthermore, we apply the brevity penalty used in BLEU (Papineni et al., 2002) to verify the accuracy of knowledge at an appropriate length, defined by the reference text for each data instance. This ensures models produce concise, non-redundant explanations.

4 Dataset Creation

The process of dataset creation, illustrated in Figure 2, involved the following steps:

STEP 1: We collected all the artwork articles from the English Wikipedia that have an infobox (about 10,000), divided them into sections, and created descriptive texts. Additionally, hyperlinked texts within the articles were extracted as entities related to the artwork. Each descriptive text is accompanied by four pieces of information: the title, the hierarchy of sections (i.e., Section, Subsection, Subsubsection), the image, and the aforementioned entities.

[2]For the formulas of each metric, see Appendix C.

LVLM	Setting	Size	BLUE	ROUGE			BertScore	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
				1	2	L		exact	partial		n=0	n=1	n=2	n=∞	
***With Title* (Language information + Visual information)**															
mPLUG-Owl2	Unseen	7B	1.16	26.8	5.9	17.1	83.3	13.3	21.1	15.6	1.61	1.38	1.35	1.29	100
LLaVA-NeXT (Vicuna-7B)	Unseen	7B	0.81	16.5	3.7	11.0	80.8	9.0	14.1	10.6	0.83	0.74	0.73	0.69	119
LLaVA-NeXT (Vicuna-13B)	Unseen	13B	1.18	17.0	4.1	10.8	80.5	11.5	16.4	13.1	1.12	1.04	1.02	0.99	133
LLaVA-NeXT (Yi-34B)	Unseen	34B	0.72	13.9	3.3	9.5	80.2	18.5	27.8	16.1	0.26	0.22	0.21	0.19	869
Qwen-VL-Chat	Unseen	7B	1.64	28.2	6.8	17.4	83.5	17.8	26.3	20.8	1.90	1.66	1.63	1.57	155
Qwen-VL-Chat (FT)	Unseen	7B	**3.96**	27.2	**10.8**	**21.4**	**84.2**	19.7	27.2	22.0	**4.86**	**4.35**	**4.23**	**4.13**	153
GPT-4-Vision	Unseen	-	2.40	**28.6**	7.6	16.3	83.3	**28.4**	**37.1**	**31.6**	3.02	3.00	2.98	3.05	264
***Without Title* (Visual information)**															
mPLUG-Owl2	Unseen	7B	0.21	23.3	3.58	15.0	82.3	4.0	10.5	4.3	0.26	0.29	0.26	0.24	91
LLaVA-NeXT (Vicuna-7B)	Unseen	7B	0.13	16.0	2.21	10.6	80.1	1.8	6.3	1.8	0.07	0.10	0.10	0.11	125
LLaVA-NeXT (Vicuna-13B)	Unseen	13B	0.17	16.6	2.35	11.0	80.8	2.1	7.1	2.2	0.07	0.08	0.08	0.07	164
LLaVA-NeXT (Yi-34B)	Unseen	34B	0.15	11.5	1.88	8.1	78.7	3.5	10.5	2.8	0.03	0.03	0.02	0.02	903
Qwen-VL-Chat	Unseen	7B	0.47	**24.8**	4.50	15.4	82.5	7.5	14.6	8.4	0.56	0.60	0.58	0.55	128
Qwen-VL-Chat (FT)	Unseen	7B	**2.07**	24.5	**7.79**	**18.6**	**83.4**	**12.9**	**19.6**	**14.7**	**2.25**	**2.03**	**2.00**	**1.96**	153
GPT-4-Vision	Unseen	-	0.10	23.1	4.43	13.2	81.9	11.6	19.0	12.3	1.18	1.35	1.37	1.34	223

Table 3: Results of LVLMs. Bold fonts indicate the best scores. Avg. Length averages generated token lengths.

STEP 2: We filtered out sections that did not contribute directly to the understanding of artwork, articles without images, and texts not specific to individual art pieces to ensure the relevance and quality of the content.

STEP 3: To prevent biases that may arise due to the notoriety of the artworks included in the LVLM's training data, we shuffled the data. First, we ranked the data using six metrics: page views, number of links, number of edits, number of references, number of language versions, and article length. We then evenly split the data into test, development, and training sets at a ratio of 1:1:8 to maintain the average ranking across these sets (Table 2). As described in §3, for the Seen set, we used training images with no overlap in reference text to prevent leakage. For the Unseen set, neither images nor reference texts are from the training set.

STEP 4: The sorted data for each set were then formatted into instructions using the templates described in Section 3.1. To diversify the training data, we prepared seven different templates inspired by Longpre et al. (2023) (see Appendix E.3).

5 Evaluation

5.1 Setup

We evaluated four models: mPLUG-Owl2 (Ye et al., 2023b), LLaVA-NeXT (Liu et al., 2024), Qwen-VL-Chat (Bai et al., 2023b), and GPT-4 Vision (OpenAI, 2023), along with an instruction-tuned version of Qwen-VL-Chat (FT), fine-tuned by our dataset with LoRA (Dettmers et al., 2022a).[3] As shown in Table 2, the data is divided based on images. In the Few-shot setting, by utilizing this data division, to prevent answer leakage in Few-shot samples, for test (Seen) evaluations, samples were randomly selected from the test (Unseen) set, and vice versa for test (Unseen) evaluations.

5.2 Results

With and Without Title Table 3 shows the results In the "With Title" setting, GPT-4-Vision achieved the highest performance in Entity Coverage and Entity F1, with Qwen-VL-Chat (FT), Qwen-VL-Chat, and LLaVA-NeXT (Yi-34B-Chat) also showing strong performance. Notably, Qwen-VL-Chat (FT) reached the highest precision in Entity Cooccurrence, showcasing its exceptional ability to accurately contextualize knowledge within generated text. This proves the superiority of our instruction-tuning dataset. Additionally, considering the average reference token length is 174 in the unseen setting, the significantly low performance of LLaVA-Next (Yi-34B-Chat) indicates excessive token lengths may result in redundant text, which is unsuitable for generating concise explanations.

In the "Without Title" setting, Qwen-VL-Chat (FT) outperformed GPT-4-Vision across all metrics, indicating that our dataset enables accurate knowledge association and generation from visual information. Comparative analysis of the models' performance in scenarios with and without titles indicated a consistent drop in performance across the board. This observation clearly shows the challenges of generating text based solely on visual inputs. All models, including advanced ones like GPT-4-Vision, heavily depend on text-based cues.

[3]Further details for the evaluation setup and results for other models are described in Appendix D and Appendix A.

[3]Since LLMs do not handle visual information, we conducted the analysis in a setting with titles.

LVLM	Setting	Size	BLUE	ROUGE			BertScore	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
				1	2	L		exact	partial		n=0	n=1	n=2	n=∞	
***With Title* (Language information + Visual information)**															
Qwen-VL-Chat	Unseen	7B	1.64	**28.2**	6.8	17.4	83.5	17.8	26.3	20.8	1.90	1.66	1.63	1.57	155
Qwen-VL-Chat One-shot	Unseen	7B	1.96	27.6	7.6	18.0	84.0	18.0	26.0	20.9	2.71	2.34	2.30	2.21	98
Qwen-VL-Chat Three-shot	Unseen	7B	2.47	27.2	8.5	18.7	**84.4**	19.3	**27.3**	**22.8**	3.65	3.14	3.05	2.97	77
Qwen-VL-Chat (FT)	Unseen	7B	**3.96**	27.2	**10.8**	**21.4**	84.2	**19.7**	27.2	22.0	**4.86**	**4.35**	**4.23**	**4.13**	153
Qwen-VL-Chat (FT) One-shot	Unseen	7B	**3.96**	26.9	10.6	21.1	84.0	**19.7**	27.0	22.0	4.75	4.20	4.02	3.97	154
Qwen-VL-Chat (FT) Three-shot	Unseen	7B	3.85	26.9	10.6	21.0	84.2	19.5	26.8	22.2	4.71	4.01	3.94	3.86	128
Qwen-VL-Chat	Seen	7B	1.69	**27.9**	6.7	17.3	83.4	16.2	24.5	19.8	1.87	1.57	1.54	1.47	153
Qwen-VL-Chat One-shot	Seen	7B	2.02	27.3	7.5	17.8	84.0	17.4	25.3	20.8	2.95	2.49	2.45	2.36	95
Qwen-VL-Chat Three-shot	Seen	7B	2.34	26.5	8.22	18.3	84.3	17.9	25.8	21.3	3.43	2.72	2.69	2.61	74
Qwen-VL-Chat (FT)	Seen	7B	**4.13**	27.6	**11.4**	**21.8**	84.5	**19.8**	**27.4**	**23.5**	**5.47**	4.43	4.30	4.19	133
Qwen-VL-Chat (FT) One-shot	Seen	7B	4.06	27.4	11.1	21.6	84.4	**19.8**	27.3	22.7	5.43	**4.45**	**4.40**	**4.30**	134
Qwen-VL-Chat (FT) Three-shot	Seen	7B	4.05	27.2	11.1	21.5	**84.6**	19.5	27.0	22.4	5.22	4.21	4.19	4.10	113
***Without Title* (Visual information)**															
Qwen-VL-Chat	Unseen	7B	0.47	**24.8**	4.50	15.4	82.5	7.5	14.6	8.4	0.56	0.60	0.58	0.55	128
Qwen-VL-Chat One-shot	Unseen	7B	0.65	23.4	4.81	15.3	83.0	8.6	15.4	9.7	1.15	1.10	1.04	1.12	87
Qwen-VL-Chat Three-shot	Unseen	7B	0.69	22.2	4.95	15.0	83.3	9.3	15.6	10.4	1.21	1.22	1.17	1.11	70
Qwen-VL-Chat (FT)	Unseen	7B	**2.07**	24.5	**7.79**	**18.6**	83.4	**12.9**	**19.6**	**14.7**	2.25	**2.03**	**2.00**	**1.96**	153
Qwen-VL-Chat (FT) One-shot	Unseen	7B	1.95	24.1	7.50	18.3	83.3	12.6	19.2	14.3	2.00	1.92	1.86	1.84	152
Qwen-VL-Chat (FT) Three-shot	Unseen	7B	2.03	24.3	7.67	18.4	**83.6**	**12.9**	**19.6**	14.6	**2.40**	2.00	1.94	1.91	131
Qwen-VL-Chat	Seen	7B	0.40	24.4	4.32	15.2	82.5	5.6	12.7	6.9	0.40	0.41	0.37	0.35	124
Qwen-VL-Chat One-shot	Seen	7B	0.53	22.5	4.45	14.8	83.0	7.2	13.9	8.6	0.72	0.72	0.70	0.66	82
Qwen-VL-Chat Three-shot	Seen	7B	0.69	22.2	4.95	15.0	83.3	9.3	15.6	10.4	1.21	1.22	1.17	1.11	68
Qwen-VL-Chat (FT)	Seen	7B	**2.09**	**24.9**	**8.00**	**18.9**	**83.8**	**12.4**	**19.4**	**15.0**	**2.19**	**1.85**	**1.82**	**1.78**	127
Qwen-VL-Chat (FT) One-shot	Seen	7B	1.99	24.4	7.72	18.5	83.6	11.5	18.7	14.0	1.89	1.55	1.51	1.48	130
Qwen-VL-Chat (FT) Three-shot	Seen	7B	2.03	24.3	7.74	18.4	**83.8**	11.6	18.5	13.9	1.89	1.49	1.45	1.42	117

Table 4: Results of Fine-tuning and Few-shot settings for LVLMs. Bold fonts indicate the best scores. Avg. Length averages generated token lengths (see Figure 4).

LLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
	exact	partial		n=0	n=1	n=2	n=∞	
***With Title* (Language information)**								
Llama2	18.5	27.3	20.8	1.04	0.88	0.82	0.81	366
Vicuna 7B	12.3	18.6	14.1	1.43	1.33	1.32	1.23	129
Vicuna 13B	19.4	28.1	23.0	2.16	1.99	1.89	1.77	209
Yi-34B-Chat	17.9	25.4	13.0	0.93	0.86	0.83	0.81	745
Qwen-Chat	7.6	11.8	8.5	0.52	0.43	0.41	0.40	106
GPT-4	**31.7**	**40.2**	**32.3**	**2.54**	**2.50**	**2.53**	**2.59**	374

Table 5: Results of LLMs (Unseen[4]). Notations are the same as Table 3.

LLMs vs. LVLMs Table 5 shows the results of explanation generation in the With Title setting without images for text-only LLMs. Notably, Table 5 illustrates that GPT-4 (OpenAI et al., 2023) achieves the highest accuracy across all metrics, demonstrating strong knowledge about artworks, closely followed by Llama2 (Touvron et al., 2023), Vicuna (Chiang et al., 2023) and Yi-34-Chat (01.AI, 2023). Conversely, Qwen-Chat (Bai et al., 2023a) is shown to perform comparatively lower. Additionally, the comparison of Tables 3 and 5 reveals the extent of text-only LLM's knowledge retention through integrated vision and language learning. It is apparent that the knowledge about artworks is compromised in other LVLMs due to the integrated learning of vision and language. On the other hand, Qwen-VL-Chat achieves a 10% performance boost in titled settings, signaling successful synthesis of vision and language knowledge.

Few-shot vs. Fine-tuning The results in Table 4 show that Fine-tuning outperforms both the pure model and Few-shot settings. While Few-shot settings show some improvement with an increasing number of shots, they do not match the performance of Fine-tuning. Considering the average token length of 174 in the reference sentences, the reduced token length in Few-shot settings suggests a focus on generating necessary terms but may result in less comprehensive explanations. In contrast, Fine-tuning allows the model to learn both specific vocabulary and the format for generating coherent explanations, leading to better performance. However, the lack of significant differences between Seen and Unseen settings in Fine-tuning indicates that effective alignment of visual and textual information (the knowledge originally held by the LLM) requires simultaneous learning of images and their descriptions.

6 Conclusion

We introduced a new task, artwork explanation generation, and its dataset and metrics to quantitatively evaluate the artistic knowledge comprehension and application. Using LVLMs, we assessed their retention and utilization of artworks knowledge from base LLMs, with or without artwork titles. Our findings indicate that while LVLMs maintain much of the artistic knowledge from their LLM counterparts, they do slightly lose some in practice. Furthermore, the challenges in generating text solely based on visual inputs clearly show a significant dependency on text-based cues.

Limitations

Our research elucidates the intricacies of integrating visual and language abilities within LVLMs, yet it encounters specific limitations that define the scope of our findings.

Data Source A principal limitation is our reliance on the diverse authorship and open editing model of Wikipedia as our data source. Variations in detail, writing style, and information density across entries may lead to inconsistencies in the dataset, potentially skewing model performance and affecting the universality of our conclusions. Additionally, we did not filter out generic entities such as "artwork" to avoid bias. However, more specific entity filtering may improve dataset relevance to artworks. Moreover, relying on Wikipedia limits our dataset to well-known artworks, omitting lesser-known but culturally significant works not featured on the platform, thereby missing a broader spectrum of artistic significance.

Human Evaluation While our current study does not include human evaluations, it is crucial to assess whether the models can provide insights beyond Wikipedia and evaluate LVLM explanations from an expert perspective for real-world applications. Another LVLM-based image explanation task, image review generation (Saito et al., 2024) actually conducts human evaluation by hiring non-expert annotators. Unlike their work, our task requires expert knowledge to judge the quality of generated explanations. Thus, due to the cost perspective, evaluating generated explanations across various genres by experts is a left problem.

Integration of Vision and Language Representations Simultaneously, our study identifies a crucial limitation in the process of integrating Vision Encoders with LLMs, particularly highlighting the models' reliance on textual cues to generate text from visual inputs. Kamigaito et al. (2023) report the same issue when predicting infoboxes, which are kinds of summaries for Wikipedia articles. This observation underscores the difficulty of retaining language knowledge during the integration, a problem we acknowledge without offering concrete solutions. This gap clearly shows the pressing need for future research to not only further investigate these issues but also to develop innovative methodologies that ensure the preservation of language knowledge amidst the integration of visual and language abilities.

Insuffcient Artwork Knowledge in LVLMs The limited improvement in entity coverage by LoRA indicates the difficulty of injecting artwork knoweldge into LVLMs. As a solution, we can consider injecting external knowledge into LVLMs. Chen et al. (2024b) introduce using knowledge graphs (KGs) as a solution. However, KGs are commonly sparse and we may need to complete them by KG completion (KGC), a task to complete missing links in KGs. Traditional KGC methods (Nickel et al., 2011; Bordes et al., 2013) are empirically (Ruffinelli et al., 2020; Ali et al., 2021) and theoretically (Kamigaito and Hayashi, 2021, 2022a,b; Feng et al., 2024) investigated in detail, and thus, these are solid whereas the pre-trained-based KGC models can outperform them (Wang et al., 2022). On the other hand, Sakai et al. (2023) point out the leakage problem of the pre-trained-based KGC models and the actual performance of them is uncertain. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) can be another solution if LVLMs can accept lengthy input (Zong et al., 2024).

Ethical Considerations

In our study, we meticulously curated our dataset derived from English Wikipedia. During the data creation phase, we individually inspected each extracted image, carefully removing those clearly unsuitable for public disclosure, ensuring no inappropriate images were included. Additionally, while English Wikipedia's editors actively eliminate unnecessarily offensive content to compile an encyclopedia, as outlined on their official pages regarding offensive material[5], bias in sources, and the use of biased or opinionated sources[6] [7], it is acknowledged that English Wikipedia allows the inclusion of biased information sources. Consequently, our dataset might also reflect the inherent biases present in the original English Wikipedia content. Note that in this work, we used an AI assistant tool, ChatGPT, for coding support.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP21K17801, JP23H03458.

[5] https://en.wikipedia.org/wiki/Wikipedia:Offensive_material

[6] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view#Bias_in_sources

[7] https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources#Biased_or_opinionated_sources

References

01.AI. 2023. Yi. https://github.com/01-ai/Yi.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2021. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research*, 22(82):1–6.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023c. Touchstone: Evaluating vision-language models by language models.

Zechen Bai, Yuta Nakashima, and Noa Garcia. 2021. Explain me the painting: Multi-topic knowledgeable art description generation.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

Eva Cetinic. 2021. Iconographic image captioning for artworks. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*, pages 502–516. Springer.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions.

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Huajun Chen. 2024b. Knowledge graphs meet multi-modal learning: A comprehensive survey.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

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

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022a. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022b. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence.

2023. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*.

Jean-Baptiste Alayrac et al. 2022. Flamingo: a visual language model for few-shot learning.

Xincan Feng, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Unified interpretation of smoothing methods for negative sampling loss functions in knowledge graph embedding.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models.

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. In *Proceedings of the European Conference in Computer Vision Workshops*.

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

Sasun Hambardzumyan, Abhinav Tuli, Levon Ghukasyan, Fariz Rahman, Hrant Topchyan, David Isayan, Mikayel Harutyunyan, Tatevik Hakobyan, Ivo Stranic, and Davit Buniatyan. 2023. Deep lake: a lakehouse for deep learning.

D. A. Hudson and C. D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, Los Alamitos, CA, USA. IEEE Computer Society.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *CVPR*.

Hidetaka Kamigaito and Katsuhiko Hayashi. 2021. Unified interpretation of softmax cross-entropy and negative sampling: With case study for knowledge graph embedding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5517–5531, Online. Association for Computational Linguistics.

Hidetaka Kamigaito and Katsuhiko Hayashi. 2022a. Comprehensive analysis of negative sampling in knowledge graph representation learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10661–10675. PMLR.

Hidetaka Kamigaito and Katsuhiko Hayashi. 2022b. Erratum to: Comprehensive analysis of negative sampling in knowledge graph representation learning.

Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2023. Table and image generation for investigating knowledge of entities in pre-trained vision and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1904–1917, Toronto, Canada. Association for Computational Linguistics.

Kento Kawaharazuka, Tatsuya Matsushima, Andrew Gambardella, Jiaxian Guo, Chris Paxton, and Andy Zeng. 2024. Real-world robot applications of foundation models: A review.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension.

Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022a. MultiSpanQA: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1260, Seattle, United States. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun. 2018. ivqa: Inverse visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8611–8619, Los Alamitos, CA, USA. IEEE Computer Society.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player?

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199, Los Alamitos, CA, USA. IEEE Computer Society.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 809–816, Madison, WI, USA. Omnipress.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook

Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

OpenAI. 2023. Gpt-4v(ision) system card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306.14824.

Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. 2020. You CAN teach an old dog new tricks! on training knowledge graph embeddings. In *International Conference on Learning Representations*.

Shigeki Saito, Kazuki Hayashi, Yusuke Ide, Yusuke Sakai, Kazuma Onishi, Toma Suzuki, Seiji Gobara, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Evaluating image review ability of vision language models.

Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2023. Does pre-trained language model actually infer unseen links in knowledge graph completion? *CoRR*, abs/2311.09109.

Alex Fang Jonathan Hayase Georgios Smyrnis Thao Nguyen Ryan Marten Mitchell Wortsman Dhruba Ghosh Jieyu Zhang Eyal Orgad Rahim Entezari Giannis Daras Sarah Pratt Vivek Ramanujan Yonatan Bitton Kalyani Marathe Stephen Mussmann Richard Vencu Mehdi Cherti Ranjay Krishna Pang Wei Koh Olga Saukh Alexander Ratner Shuran Song Hannaneh Hajishirzi Ali Farhadi Romain Beaumont Sewoong Oh Alex Dimakis Jenia Jitsev Yair Carmon Vaishaal Shankar Ludwig Schmidt Samir Yitzhak Gadre, Gabriel Ilharco. 2023. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

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.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. *A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge*, pages 146–162.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8309–8318.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022. SimKGC: Simple contrastive knowledge graph completion with pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4281–4294, Dublin, Ireland. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*.

Tomas Yago, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. 2016. Large-scale training of shadow detectors with noisily-annotated shadow examples. volume 9910, pages 816–832.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, and Fei Huang. 2023a. mplug-owl: Modularization empowers large language models with multimodality.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration.

Licheng Yu, Patric Poirson, Shan Yang, Alexander Berg, and Tamara Berg. 2016. Modeling context in referring expressions.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022a. Opt: Open pre-trained transformer language models.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yao Zhang, Haokun Chen, Ahmed Frikha, Yezi Yang, Denis Krompass, Gengyuan Zhang, Jindong Gu, and Volker Tresp. 2022b. Cl-crossvqa: A continual learning benchmark for cross-domain visual question answering.

Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. 2024. Vl-icl bench: The devil in the details of benchmarking multimodal in-context learning. *arXiv preprint arXiv:2403.13164*.

A Supplemental Results

A.1 Detailed Evaluation of LVLMs in 'Seen' Data Settings

Table 8 presents the results of Language-Vision Learning Models (LVLMs) including 'seen' settings, with bold type highlighting the highest score for each metric within each group. In this study, we assessed the generalizability of data and the precision of models fine-tuned on 'seen' and 'unseen' data during their training phase to ascertain if the fine-tuning process enhanced the models' accuracy for images encountered during training. Despite the images being part of the training dataset, with sections meticulously segregated to prevent data leakage, our validation revealed no significant differences in accuracy between 'seen' and 'unseen' settings. This finding confirms the general applicability of the data and suggests that simply viewing images, without integrating them with relevant contextual knowledge, does not inherently contribute to accuracy improvement. This highlights the importance of a holistic learning approach where images are paired with pertinent information to truly boost the performance of the models.

Furthermore it is generally impractical to create datasets that combine images corresponding to the vast amounts of text data seen during the training of LLMs and to acquire these through additional integrated learning. Additionally, during the integrated learning process from LLM to LVLM, the focus is on learning pairs of individual images and their descriptions. To develop the ability to individually recognize knowledge objects and explain them based on that recognition, as well as to understand the relationships between objects and generate comprehensive explanations, it is considered necessary to use enhancement methods such as RAG and new integrated learning techniques for LVLMs.

A.2 Extended Analysis of Additional LVLMs

In our research, we expanded our experimental investigation beyond the models outlined in the primary section to include Blip2 (Li et al., 2023b), mPLUG_Owl (Ye et al., 2023a), LLaVA-NeXT (Mistral) (Liu et al., 2024), LLaVA-1.5 (Liu et al., 2023a,b), InstructBlip (Dai et al., 2023), and Yi-6B (01.AI, 2023), integrating image and language in a manner similar to the initially described models. Utilizing the same experimental framework as the initial tests, we conducted an thorough assessment. The results, as outlined in Table 9, revealed that these additional models did not exceed the accuracy levels of those featured in the main analysis (refer to Section 5). Additionally, a comparative examination of configurations with and without titles showed a uniform decline in efficacy, emphasizing the difficulty of deriving knowledge and translating it into explanatory text generation based purely on image data.

A.3 Detailed Performance Metrics for Base LLMs with Title Context

Table 10 presents the results of an evaluation involving the base LLM models of the Language-Vision Learning Models (LVLMs) discussed in Tables 3 and 9. This evaluation additionally included tests on base models such as FLAN-T5-XL(Chung et al., 2022), FLAN-T5-XXL, OPT(Zhang et al., 2022a), LLaMA(Touvron et al., 2023) Mistral(Jiang et al., 2023), and Yi-6B, which were not featured in the main analysis. Since Language Models (LMs) are incapable of processing image information, the evaluation was confined to the 'With Title' setting that incorporates textual information. Within this context, GPT-4 showcased superior performance across all tested configurations, with Mistral, Vicuna-13B, and LlaMA2 also demonstrating strong results.

Consistent with the data presented in Table 3, the base model for LLaVA-NeXT (Yi-34B) yielded output sequences with excessively token lengths compared to its counterparts, mirroring the behavior of its LVLM version. This tendency for producing longer output is illustrated when compared with other models (as depicted in Figure 3). Furthermore, when examining the accuracy of the LVLMs tested in Table 9 alongside the base models in relation to our task proposal, there is a discernible decline in precision across nearly all models. Qwen is the exception, which highlights the nuanced challenges in effectively merging image and textual data. This complexity stands as a pivotal challenge for the evolution of sophisticated LVLMs.

B Title generation

In our task, the titles of artworks are a crucial element of knowledge related to the artworks. To maintain the integrity of the analysis between the settings with and without titles setting, we intentionally omitted titles from entity recognition. However, we recognized the need to understand the performance of models in generating titles of artworks based solely on visual information. Therefore, We conducted an additional experiment in which we presented the models with the prompt **"Please answer the title of this artwork"** along with 963 images from the "Unseen" test set and evaluated the accuracy of title generation under two settings: Exact and Partial. Tables 11, 12 and 13 display the accuracy results of the main models and those from additional experiments, respectively.

The results showed that GPT-4-Vision achieved the highest performance with an exact match setting at 8.97%, followed by Qwen-VL-Chat (FT) and Qwen-VL-Chat with good performances. Other models scored 2% or less, highlighting the difficulty of generating titles. Additionally, none of the LLaVA-NeXT models were able to correctly generate a single title.

Furthermore, Table 14 shows the actual artwork titles generated by the top five models with the best accuracy in the exact match setting. The "Rank" in the table is used to distribute the dataset evenly at the time of its creation (refer to Section 3), between famous and less famous paintings, to prevent bias. From the table, we can infer that a higher proportion of famous artworks with higher ranks were generated, indicating that the models have a better grasp of more famous artworks.

C Evaluation Metrics Formulation

This section elaborates on the evaluation metrics proposed in Section 3.2 using mathematical expressions. An explanation consisting of n sentences generated by the model is denoted as $G = \{g_1, \cdots, g_n\}$, and a reference explanation consisting of m sentences is denoted as $R = \{r_1, \cdots, r_m\}$. The function Entity$(\cdot)$ is defined to extract entities contained in the input text. The notation $|G|$ represents the total number of tokens in the generated explanation, and $|R|$ represents the total number of tokens in the reference explanation.

Entity Coverage (EC) is calculated as follows:

$$EC(G, R) = Cov(G, R) \tag{1}$$

Here, $Cov(G, R)$ is a function returning the proportion of entities in R that are covered by G. For partial matches, the Lowest Common Subsequence (LCS) is employed to calculate the longest matching length ratio in the generated explanation relative to the length of the reference entity.

Entity F1 ($\mathbf{EF}_1$) is computed as follows:

$$EF_1 = \frac{2 \times P \times R}{P + R} \tag{2}$$

$$P = \frac{\sum_{e_i \in Entity(G)} \text{Count}_{\text{clip}}(e_i, G, R)}{\sum_{e_j \in Entity(G)} \#(e_j, G)} \tag{3}$$

$$R = \frac{\sum_{e_i \in Entity(R)} \text{Count}_{\text{clip}}(e_i, G, R)}{\sum_{e_j \in Entity(R)} \#(e_j, R)}, \tag{4}$$

where $\#(e_j, G)$, $\#(e_j, R)$ are functions that count the occurrences of entity e_j in G and R respectively, and $\text{Count}_{\text{clip}}(e_i, G, R)$ returns the lesser frequency of occurrence of e_i in either G or R.

Entity Cooccurrence (ECooc) is calculated using BP from equation (6) as follows:

$$\begin{aligned} &ECooc(G, R) \\ =&BP(G, R) \times Cov(Co(G), Co(R)), \end{aligned} \tag{5}$$

where $BP(G, R)$ is given by:

$$BP(G, R) = \exp(\max(0.0, \frac{|G|}{|R|} - 1)) \tag{6}$$

and function $Co(\cdot)$ returns pairs of co-occurring entities within a context window comprising a sentence and its adjacent n sentences. Sentence segmentation was performed using the nltk sentence splitter for this purpose.[8]

D Details of experimental setting

D.1 LVLM details

Model	Base Model	HuggingFace Name/OpenAI API
BLIP2 (OPT)	OPT	Salesforce/blip2-opt-6.7b
BLIP2 (FLAN-T5-XL)	FLAN-T5-XL	Salesforce/blip2-flan-t5-xl
BLIP2 (FLAN-T5-XXL)	FLAN-T5-XXL	Salesforce/blip2-flan-t5-xxl
InstructBLIP (FLAN-T5-XL)	FLAN-T5-XL	Salesforce/instructblip-flan-t5-xl
InstructBLIP (FLAN-T5-XXL)	FLAN-T5-XXL	Salesforce/instructblip-flan-t5-xxl
InstructBLIP (Vicuna-7B)	Vicuna-7B	Salesforce/instructblip-vicuna-7b
InstructBLIP (Vicuna-13B)	Vicuna-13B	Salesforce/instructblip-vicuna-13b
Yi-VL-6B	Yi-6B-Chat	01-ai/Yi-VL-6B
mPLUG-Owl	LLaMA	MAGAer13/mplug-owl-llama-7b
mPLUG-Owl2	LLaMA2-7B	MAGAer13/mplug-owl2-llama2-7b
LLaVA-1.5	Vicuna-13B	liuhaotian/llava-v1.5-13b
LLaVA-NeXT (Vicuna-7B)	Vicuna-7B	liuhaotian/llava-v1.6-vicuna-7b
LLaVA-NeXT (Vicuna-13B)	Vicuna-13B	liuhaotian/llava-v1.6-vicuna-13b
LLaVA-Next (Mistral)	Mistral	liuhaotian/llava-v1.6-mistral-7b
LLaVA-NeXT (Yi-34B)	Yi-34B	liuhaotian/llava-v1.6-34b
Qwen-VL-Chat	Qwen	Qwen/Qwen-VL-Chat
GPT-4-Vision	-	gpt-4-1106-vision-preview

[8]Sentence segmentation was performed using the NLTK sentence splitter.

D.2 LLM details

Model	HuggingFace Name
FLAN-T5-XL	google/flan-t5-xl
FLAN-T5-XXL	google/flan-t5-xxl
OPT	facebook/opt-6.7b
LLaMA	openlm-research/open_llama_7b
LLaMA2	meta-llama/Llama-2-7b
Mistral	mistralai/Mistral-7B-Instruct-v0.2
Vicuna-7B	lmsys/vicuna-7b-v1.5
Vicuna-13B	lmsys/vicuna-13b-v1.5
Qwen-Chat	Qwen/Qwen-7B-Chat
Yi-6B	01-ai/Yi-6B
Yi-34B	01-ai/Yi-34B
GPT-4	gpt-4-1106-preview

D.3 Fine tunning and Inference setting

Hyper Parameter	Value
torch_dtype	bfloat16
seed	42
max length	2048
warmup ratio	0.01
learning rate	1e-5
batch size	4
epoch	1
lora r	64
lora alpha	16
lora dropout	0.05
lora target modules	c_attn, attn.c_proj, w1, w2

Table 6: The hyper-parameters used in the experiment, and others, were set to default settings. The implementation used Transformers (Wolf et al., 2020) and bitsandbytes (Dettmers et al., 2022b).

In this study, to ensure a fair comparison of performance across multiple models, all experiments were conducted on a single NVIDIA RTX 6000 Ada GPU, with 8-bit quantization utilized for model generation. However, due to resource constraints, LLaVA-NeXT (Yi-34B-Chat) model was loaded and inferred in 4-bit mode. To standardize the length of tokens generated across all models, the maximum token length was set to 1024. The same settings were applied to each model for performance comparison purposes.

D.4 Training Datasets

Table 16 lists the datasets employed to train the models addressed in this study.

E Details of our created dataset

E.1 Dataset section distribution

Table 7 provides a comprehensive breakdown of various types of sections within the dataset, along with their frequency counts. In designing the test set for the "seen" setting, we meticulously considered the distribution of these sections. Through an analysis of the frequency of each section type, we managed to evenly split the data. This strategic approach ensured that the test set was constructed with a balanced representation of each section type, aiming for a more equitable and thorough evaluation process. Due to this methodology, the division of the test set into "seen" and "unseen" portions was based on the distribution of section types, rather than the number of images. Consequently, the number of images in the "seen" and "unseen" parts of the test set may not be equal (refer to Table 2). This was a deliberate choice to prioritize a balanced representation of section types over an equal count of images, enhancing the relevance and fairness of the evaluation process.

E.2 Omitted sections

The following sections have been omitted from this document:

- References
- See also
- External links
- Sources
- Further reading
- Bibliography
- Gallery
- Footnotes
- Notes
- References Sources
- Bibliography (In Spanish)
- Bibliography (In Italian)
- Bibliography (In German)
- Bibliography (In French)
- Images
- Links
- List
- Notes and references
- List by location

These sections were deemed unsuitable for the task of generating descriptions of artwork in this study and were therefore removed.

E.3 Train Templates

As shown in Table 15, to ensure diversity in training, we utilized seven templates to construct the instruction-based training set. We initially created 49 templates by combining seven base sentences with seven verbs such as explore, explain, and discuss. During experimental evaluations, the models were tested with these 49 templates. We adopted the top seven templates that resulted in the highest accuracy and best adherence to instructions by the models.

E.4 Train Dataset Example

As shown in Figure 5 and 6, we adopted the format for fine-tuning Qwen (Bai et al., 2023a) and modified the template presented in E.3 into the form of figures. This format was used for model training and dataset publication.

E.5 Entity Distribution

Figures 7 and 8 present the entity distribution within our datasets. The minimal difference in data distribution between seen and unseen cases suggests that the partitioning method described in Step 3 of Section 4 is effective.

F License

In our study we created a dataset from Wikipedia articles of artworks. The each image is available under the Creative Commons License (CC) or other licenses. Specific license information for each image can be found on the Wikipedia page or the image description page for that image. The images in this study are used under the terms of these licenses, and links to the images are provided in the datasets we publish so that users can download the images directly. The images themselves are not directly published. Therefore, our data does not infringe upon the licenses.

Type	**Frequency**
Abstract	9632
Description	2747
History	1869
Background	666
Provenance	517
Reception	346
Description History	341
Analysis	337
Painting	218
Artist	189
Historical Information	187
Composition	168
Subject	138
Legacy	127
Exhibitions	115
Interpretation	110
Condition	97
In Popular Culture	94
Information	84
Design	83
Style	78
Influence	68
Creation	65
Description Style	63
Related Works	63
Acquisition	60
Context	59
Versions	51
Other Versions	51
Literature	50
Symbolism	50
The Painting	50
Attribution	50
Details	46
Notes References	45
Exhibition History	41
Location	40
Interpretations	40
Critical Reception	39
Historical Context	39
Iconography	38
Subject Matter	37
Influences	37
Exhibition	37
Commission	36
Overview	34
Analysis Description	34
Citations	33
Painting Materials	32
Controversy	32
Restoration	32

Table 7: Frequency count of data types in the dataset.

LVLM	Setting	Size	BLUE	ROUGE			BertScore	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
				1	2	L		exact	partial		n=0	n=1	n=2	n=∞	
***With Title* (Language information + Visual information)**															
mPLUG-Owl2	Unseen	7B	1.16	26.8	5.9	17.1	83.3	13.3	21.1	15.6	1.61	1.38	1.35	1.29	100
LLaVA-NeXT (Vicuna-7B)	Unseen	7B	0.81	16.5	3.7	11.0	80.8	9.0	14.1	10.6	0.83	0.74	0.73	0.69	119
LLaVA-NeXT (Vicuna-13B)	Unseen	13B	1.18	17.0	4.1	10.8	80.5	11.5	16.4	13.1	1.12	1.04	1.02	0.99	133
LLaVA-NeXT (Yi-34B)	Unseen	34B	0.72	13.9	3.3	9.5	80.2	18.5	27.8	16.1	0.26	0.22	0.21	0.19	869
Qwen-VL-Chat	Unseen	7B	1.64	28.2	6.8	17.4	83.5	17.8	26.3	20.8	1.90	1.66	1.63	1.57	155
Qwen-VL-Chat (FT)	Unseen	7B	**3.96**	27.2	**10.8**	**21.4**	**84.2**	19.7	27.2	22.0	**4.86**	**4.35**	**4.23**	**4.13**	153
GPT-4-Vision	Unseen	-	2.40	**28.6**	7.6	16.3	83.3	**28.4**	**37.1**	**31.6**	3.02	3.00	2.98	3.05	264
mPLUG-Owl2	Seen	7B	1.14	26.6	5.9	17.0	83.3	12.5	20.3	15.1	1.54	1.29	1.24	1.17	94
LLaVA-NeXT (Vicuna-7B)	Seen	7B	0.78	16.5	3.5	10.6	80.7	7.9	13.0	9.4	0.74	0.66	0.63	0.59	114
LLaVA-NeXT (Vicuna-13B)	Seen	13B	1.14	17.0	4.0	10.8	80.5	10.3	15.5	12.4	1.32	1.08	1.01	0.96	127
LLaVA-NeXT (Yi-34B)	Seen	34B	0.73	13.7	3.2	9.4	80.1	17.4	26.7	15.4	0.26	0.24	0.22	0.21	872
Qwen-VL-Chat	Seen	7B	1.69	**27.9**	6.7	17.3	83.4	16.2	24.5	19.8	1.87	1.57	1.54	1.47	153
Qwen-VL-Chat (FT)	Seen	7B	**4.13**	27.6	**11.4**	**21.8**	**84.5**	**19.8**	**27.4**	**23.5**	**5.47**	**4.43**	**4.30**	**4.19**	133
GPT-4-Vision	Seen	-	2.32	**28.3**	7.4	16.2	83.2	**26.4**	**34.9**	**29.7**	2.82	2.71	2.67	2.63	254
***Without Title* (Visual information)**															
mPLUG-Owl2	Unseen	7B	0.21	23.3	3.58	15.0	82.3	4.0	10.5	4.3	0.26	0.29	0.26	0.24	91
LLaVA-NeXT (Vicuna-7B)	Unseen	7B	0.13	16.0	2.21	10.6	80.1	1.8	6.3	1.8	0.07	0.10	0.10	0.11	125
LLaVA-NeXT (Vicuna-13B)	Unseen	13B	0.17	16.6	2.35	11.0	80.8	2.1	7.1	2.2	0.07	0.08	0.08	0.07	164
LLaVA-NeXT (Yi-34B)	Unseen	34B	0.15	11.5	1.88	8.1	78.7	3.5	10.5	2.8	0.03	0.03	0.02	0.02	903
Qwen-VL-Chat	Unseen	7B	0.47	**24.8**	4.50	15.4	82.5	7.5	14.6	8.4	0.56	0.60	0.58	0.55	128
Qwen-VL-Chat (FT)	Unseen	7B	**2.07**	24.5	**7.79**	**18.6**	**83.4**	**12.9**	**19.6**	**14.7**	**2.25**	**2.03**	**2.00**	**1.96**	153
GPT-4-Vision	Unseen	-	0.10	23.1	4.43	13.2	81.9	11.6	19.0	12.3	1.18	1.35	1.37	1.34	223
mPLUG-Owl2	Seen	7B	0.14	22.6	3.37	14.6	82.2	2.9	9.2	3.2	0.19	0.14	0.13	0.12	86
LLaVA-NeXT (Vicuna-7B)	Seen	7B	0.11	15.4	1.95	10.2	80.0	1.0	5.6	1.2	0.05	0.04	0.06	0.06	123
LLaVA-NeXT (Vicuna-13B)	Seen	13B	0.11	16.0	2.10	10.7	80.7	1.2	6.0	1.4	0.03	0.03	0.03	0.03	154
LLaVA-NeXT (Yi-34B)	Seen	34B	0.10	11.1	1.71	7.9	78.6	2.1	9.2	1.9	0.01	0.01	0.01	0.01	909
Qwen-VL-Chat	Seen	7B	0.40	24.4	4.32	15.2	82.5	5.6	12.7	6.9	0.40	0.41	0.37	0.35	124
Qwen-VL-Chat (FT)	Seen	7B	**2.09**	**24.9**	**8.00**	**18.9**	**83.8**	**12.4**	**19.4**	**15.0**	**2.19**	**1.85**	**1.82**	**1.78**	127
GPT-4-Vision	Seen	-	0.74	22.4	4.14	12.8	81.8	9.3	16.7	10.5	0.91	0.91	0.86	0.84	212

Table 8: Results of LVLMs including 'seen' settings. Notations are the same as Table 3.

LVLM	Setting	Size	BLUE	ROUGE			BertScore	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
				1	2	L		exact	partial		n=0	n=1	n=2	n=∞	
With Title **(Language information + Visual information)**															
BLIP2 (OPT)	Unseen	6.7B	0.00	0.1	0.0	0.1	76.4	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.01
BLIP2 (FLAN-T5-XL)	Unseen	3B	0.00	9.7	2.8	8.3	80.6	5.2	8.5	1.4	0.05	0.03	0.03	0.03	20
BLIP2 (FLAN-T5-XXL)	Unseen	11B	0.01	2.8	0.5	2.6	76.5	0.7	2.4	0.5	0.01	0.00	0.00	0.00	21
mPLUG-Owl	Unseen	7B	0.17	15.0	2.4	10.1	81.8	4.3	8.6	4.7	0.35	0.38	0.40	0.37	12
LLaVA-1.5	Unseen	13B	1.61	20.8	5.2	13.2	81.5	13.4	19.4	15.8	1.56	1.34	1.33	1.26	139
LLaVA-NeXT (Mistral)	Unseen	7B	1.32	24.1	5.7	15.9	82.4	12.3	19.6	14.9	1.44	1.18	1.15	1.06	140
InstructBLIP (FLAN-T5-XL)	Unseen	3B	0.70	16.9	5.2	13.0	83.2	8.5	13.8	6.6	0.80	0.62	0.59	0.56	28
InstructBLIP (FLAN-T5-XXL)	Unseen	11B	1.00	16.4	4.6	12.0	81.7	8.6	13.8	9.3	1.00	0.75	0.73	0.71	54
InstructBLIP (Vicuna-7B)	Unseen	7B	1.44	23.5	6.2	15.7	83.3	12.6	19.2	14.2	1.79	1.50	1.44	1.38	58
InstructBLIP (Vicuna-13B)	Unseen	13B	1.11	25.9	6.2	17.2	83.6	11.8	18.8	13.7	1.42	1.19	1.16	1.09	50
Yi-VL-6B	Unseen	6B	1.07	26.2	5.7	16.6	82.9	12.9	20.8	15.1	1.37	1.24	1.27	1.21	147
Qwen-VL-Chat	Unseen	7B	1.64	28.2	6.8	17.4	83.5	17.8	26.3	20.8	1.90	1.66	1.63	1.57	155
Qwen-VL-Chat (FT)	Unseen	7B	**3.96**	27.2	**10.8**	**21.4**	**84.2**	19.7	27.2	22.0	**4.86**	**4.35**	**4.23**	**4.13**	153
GPT-4-Vision	Unseen	-	2.40	**28.6**	7.6	16.3	83.3	**28.4**	**37.1**	**31.6**	3.02	3.00	2.98	3.05	264
BLIP2 (OPT)	Seen	6.7B	0.00	2.0	0.0	1.2	77.5	0.0	1.8	0.0	0.00	0.00	0.00	0.00	0.01
BLIP2 (FLAN-T5-XL)	Seen	3B	0.01	9.9	3.0	8.5	80.7	5.2	8.3	1.7	0.07	0.03	0.03	0.03	17
BLIP2 (FLAN-T5-XXL)	Seen	11B	0.01	2.9	0.5	2.7	76.5	0.9	2.6	0.6	0.04	0.03	0.03	0.03	21
mPLUG-Owl	Seen	7B	0.14	15.4	2.4	10.3	81.9	4.5	9.3	4.8	0.37	0.29	0.28	0.26	13
LLaVA-1.5	Seen	13B	1.69	20.7	5.3	13.1	81.5	12.5	18.4	15.0	1.85	1.37	1.34	1.30	128
LLaVA-NeXT (Mistral)	Seen	7B	1.41	24.1	5.6	16.0	82.3	11.6	19.1	14.4	1.49	1.16	1.06	1.01	145
InstructBLIP (FLAN-T5-XL)	Seen	3B	0.78	16.9	5.2	13.0	83.2	8.5	14.0	7.1	0.92	0.69	0.66	0.63	29
InstructBLIP (FLAN-T5-XXL)	Seen	11B	0.10	16.6	4.7	12.2	81.8	8.7	14.1	9.3	1.11	0.90	0.87	0.84	54
InstructBLIP (Vicuna-7B)	Seen	7B	1.53	23.9	6.3	15.8	83.3	12.4	19.5	14.3	1.77	1.47	1.42	1.37	62
InstructBLIP (Vicuna-13B)	Seen	13B	1.11	25.5	6.1	16.9	83.5	10.2	17.3	12.5	1.26	1.08	1.01	0.97	51
Yi-VL-6B	Seen	6B	1.00	25.8	5.5	16.3	82.7	11.5	19.9	13.6	1.00	0.80	0.78	0.75	149
Qwen-VL-Chat	Seen	7B	1.69	**27.9**	6.7	17.3	83.4	16.2	24.5	19.8	1.87	1.57	1.54	1.47	153
Qwen-VL-Chat (FT)	Seen	7B	**4.13**	27.6	**11.4**	**21.8**	**84.5**	**19.8**	**27.4**	**23.5**	**5.47**	**4.43**	**4.30**	**4.19**	133
GPT-4-Vision	Seen	-	2.32	**28.3**	7.4	16.2	83.2	**26.4**	**34.9**	**29.7**	2.82	2.71	2.67	2.63	254
Without Title **(Visual information)**															
BLIP2 (OPT)	Unseen	6.7B	0.00	4.1	0.00	4.1	79.8	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.01
BLIP2 (FLAN-T5-XL)	Unseen	3B	0.01	8.9	1.47	7.5	81.2	2.1	5.0	1.1	0.01	0.00	0.00	0.00	15
BLIP2 (FLAN-T5-XXL)	Unseen	11B	0.00	2.5	0.16	2.4	75.8	0.6	1.7	0.2	0.00	0.00	0.00	0.00	18
mPLUG-Owl	Unseen	7B	0.14	18.1	2.59	11.9	82.1	2.2	7.2	2.4	0.13	0.10	0.08	0.08	21
LLaVA-1.5	Unseen	13B	0.21	17.8	2.70	11.7	81.4	2.7	7.9	2.6	0.11	0.15	0.15	0.15	158
LLaVA-NeXT (Mistral)	Unseen	7B	0.16	21.1	2.77	14.1	81.3	2.3	8.0	2.3	0.08	0.11	0.12	0.12	132
InstructBLIP (FLAN-T5-XL)	Unseen	3B	0.08	13.0	2.17	10.0	82.4	2.7	6.6	2.3	0.13	0.07	0.08	0.07	28
InstructBLIP (FLAN-T5-XXL)	Unseen	11B	0.16	12.5	2.11	9.3	81.1	3.0	6.9	2.7	0.16	0.13	0.11	0.11	41
InstructBLIP (Vicuna-7B)	Unseen	7B	0.49	22.9	4.47	15.2	82.9	6.4	12.9	7.1	0.55	0.58	0.56	0.49	83
InstructBLIP (Vicuna-13B)	Unseen	13B	0.39	23.5	4.31	15.8	82.8	4.8	11.5	5.2	0.37	0.33	0.31	0.28	85
Yi-VL-6B	Unseen	6B	0.37	23.4	4.08	15.1	82.0	5.4	12.2	5.7	0.35	0.36	0.35	0.34	158
Qwen-VL-Chat	Unseen	7B	0.47	**24.8**	4.50	15.4	82.5	7.5	14.6	8.4	0.56	0.60	0.58	0.55	128
Qwen-VL-Chat (FT)	Unseen	7B	**2.07**	24.5	**7.79**	**18.6**	**83.4**	**12.9**	**19.6**	**14.7**	**2.25**	**2.03**	**2.00**	**1.96**	153
GPT-4-Vision	Unseen	-	0.10	23.1	4.43	13.2	81.9	11.6	19.0	12.3	1.18	1.35	1.37	1.34	223
BLIP2 (OPT)	Seen	6.7B	0.00	2.3	0.00	2.3	78.4	0.0	2.1	0.0	0.00	0.00	0.00	0.00	0.03
BLIP2 (FLAN-T5-XL)	Seen	3B	0.00	9.0	1.50	7.6	81.4	1.7	4.5	1.0	0.01	0.01	0.01	0.01	13
BLIP2 (FLAN-T5-XXL)	Seen	11B	0.00	2.6	0.16	2.5	75.7	0.4	1.6	0.2	0.00	0.00	0.00	0.00	18
mPLUG-Owl	Seen	7B	0.08	18.4	2.64	12.1	82.1	1.9	6.9	2.5	0.08	0.05	0.04	0.04	23
LLaVA-1.5	Seen	13B	0.13	17.7	2.55	11.6	81.3	1.3	6.4	1.4	0.07	0.05	0.05	0.04	154
LLaVA-NeXT (Mistral)	Seen	7B	0.08	20.7	2.50	13.9	81.3	1.3	7.0	1.4	0.04	0.04	0.04	0.03	125
InstructBLIP (FLAN-T5-XL)	Seen	3B	0.05	12.5	1.99	9.6	82.4	1.9	5.9	1.9	0.04	0.06	0.06	0.06	26
InstructBLIP (FLAN-T5-XXL)	Seen	11B	0.10	12.3	1.95	9.1	81.1	2.3	6.3	2.2	0.08	0.08	0.07	0.07	37
InstructBLIP (Vicuna-7B)	Seen	7B	0.43	22.7	4.31	15.1	83.0	4.9	11.4	5.8	0.36	0.30	0.29	0.27	82
InstructBLIP (Vicuna-13B)	Seen	13B	0.37	23.3	4.27	15.7	82.7	3.3	10.0	4.0	0.17	0.16	0.16	0.15	85
Yi-VL-6B	Seen	6B	0.33	23.0	3.86	14.8	81.9	4.1	11.2	4.7	0.19	0.16	0.15	0.14	162
Qwen-VL-Chat	Seen	7B	0.40	24.4	4.32	15.2	82.5	5.6	12.7	6.9	0.40	0.41	0.37	0.35	124
Qwen-VL-Chat (FT)	Seen	7B	**2.09**	**24.9**	**8.00**	**18.9**	**83.8**	**12.4**	**19.4**	**15.0**	**2.19**	**1.85**	**1.82**	**1.78**	127
GPT-4-Vision	Seen	-	0.74	22.4	4.14	12.8	81.8	9.3	16.7	10.5	0.91	0.91	0.86	0.84	212

Table 9: Comprehensive Results of Secondary (LVLMs). This includes models not highlighted in the main findings, with the gray lines representing the three models that achieved the best performance in the main evaluation. Bold type signifies the highest scores for each metric within their respective groups.

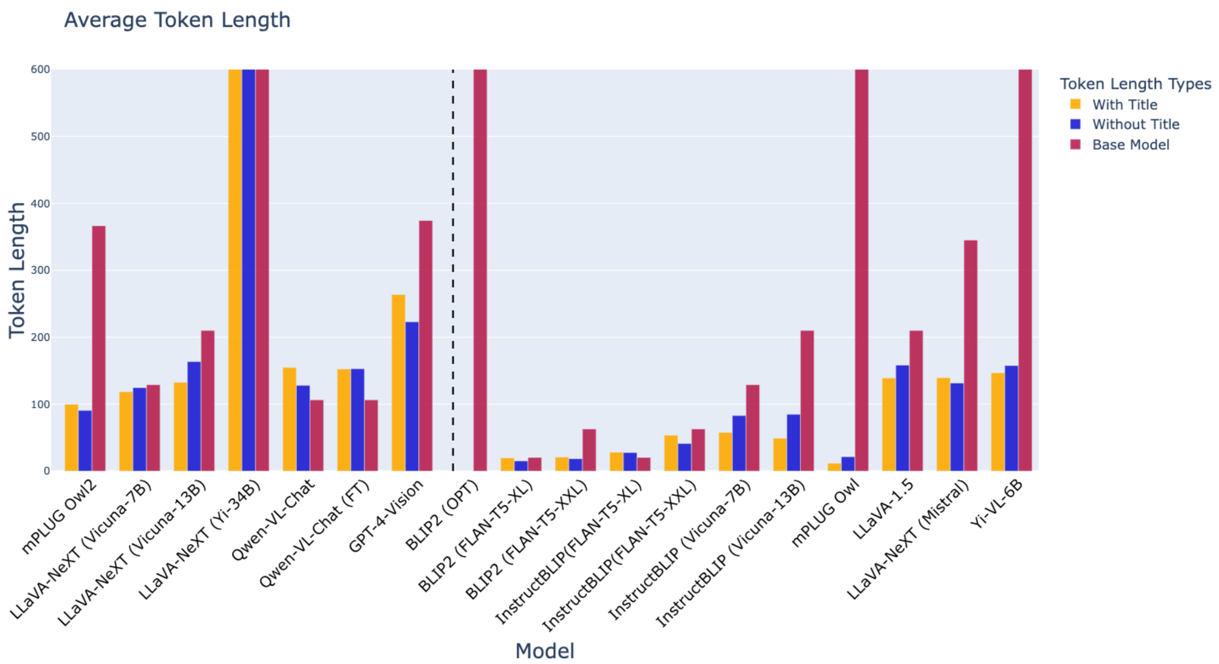

Figure 3: Average token lengths for 18 evaluated LVLMs on an unseen set, where yellow represents the 'With Title' setting, blue indicates the 'Without Title' setting, and red signifies the average token length for the base language model of the LVLM with titles. The length of the unseen reference sentence is 174 tokens.

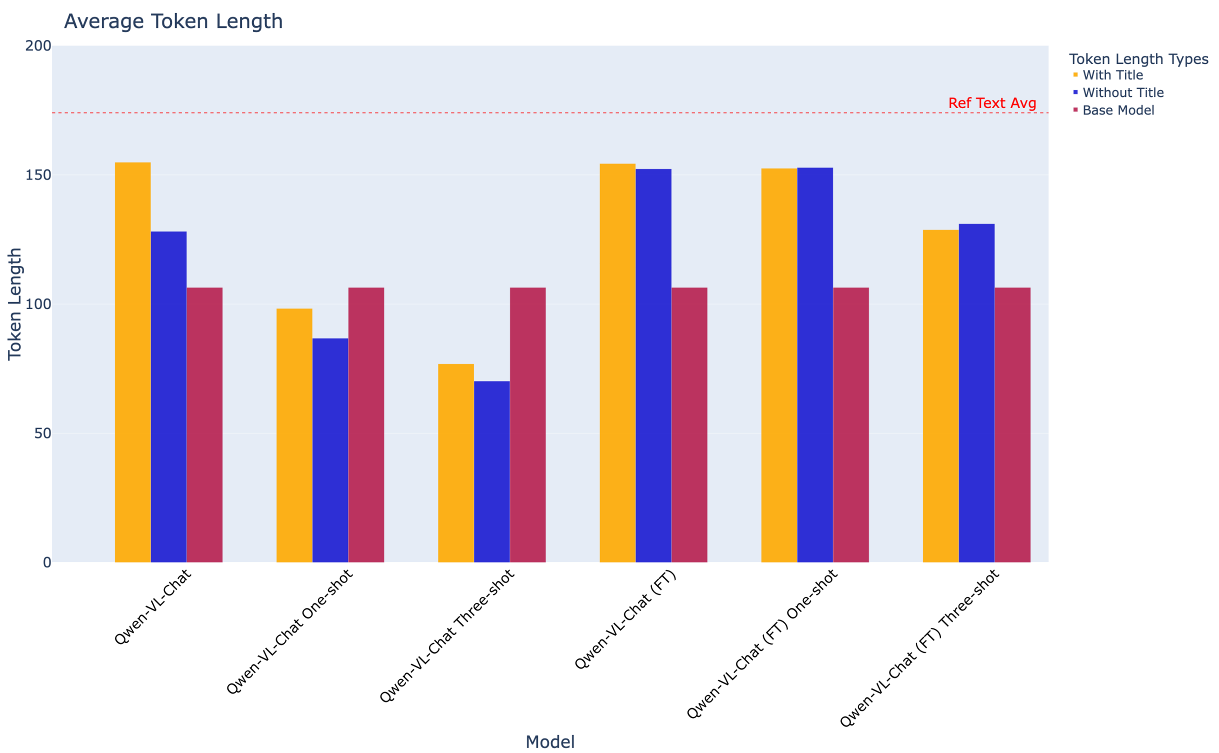

Figure 4: Average token lengths for Qwen's Few-shot and Fine-tuning settings on an unseen set, where yellow represents the 'With Title' setting, blue indicates the 'Without Title' setting, and red signifies the average token length for the base language model of the LVLM with titles. The length of the unseen reference sentence is 174 tokens.

LVLM	Setting	Size	BLUE	ROUGE			BertScore	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
				1	2	L		exact	partial		n=0	n=1	n=2	n=∞	
With Title **(Language information + Visual information)**															
FLAN-T5-XL	Unseen	3B	0.66	15.4	6.23	13.1	83.6	10.2	15.4	10.6	1.36	0.88	0.84	0.83	20
FLAN-T5-XXL	Unseen	11B	0.00	2.0	0.09	1.8	76.2	3.3	2.2	0.3	0.00	0.00	0.00	0.00	63
OPT	Unseen	6.7B	0.34	8.3	1.60	7.3	76.8	12.0	18.9	8.4	0.15	0.12	0.12	0.11	872
LlaMA	Unseen	7B	0.48	9.4	1.99	8.1	77.7	16.4	23.7	11.3	0.15	0.14	0.13	0.11	876
LlaMA2	Unseen	7B	1.81	24.0	5.92	14.9	82.4	18.5	27.3	20.8	1.04	0.88	0.82	0.81	366
Mistral	Unseen	7B	1.82	25.1	6.41	15.2	82.7	21.8	31.2	23.4	1.33	1.30	1.27	1.25	345
Vicuna-7B	Unseen	7B	1.14	20.9	4.87	13.1	82.7	12.3	18.6	14.1	1.43	1.33	1.32	1.23	129
Vicuna-13B	Unseen	13B	2.35	28.4	7.34	17.7	83.4	19.4	28.1	23.0	2.16	1.99	1.89	1.77	210
Qwen-Chat	Unseen	7B	0.60	12.0	2.50	7.4	79.5	7.6	11.8	8.5	0.52	0.43	0.41	0.40	106
Yi-6B-Chat	Unseen	6B	0.93	14.0	3.55	10.9	79.3	14.2	21.4	11.9	0.55	0.50	0.48	0.46	717
Yi-34B-Chat	Unseen	34B	1.00	13.1	3.50	10.4	79.1	17.9	25.4	12.9	0.93	0.86	0.83	0.81	745
GPT-4	Unseen	-	2.20	26.2	7.00	14.9	82.5	31.7	40.2	32.3	2.54	2.50	2.53	2.59	374
FLAN-T5-XL	Seen	3B	0.67	15.1	6.30	12.9	83.4	9.0	14.5	9.5	1.34	0.95	0.85	0.81	22
FLAN-T5-XXL	Seen	11B	0.01	8.9	1.48	7.5	81.2	2.1	5.0	1.1	0.01	0.00	0.00	0.00	66
OPT	Seen	6.7B	0.35	8.3	1.63	7.2	76.8	11.4	18.4	9.0	0.08	0.06	0.05	0.05	877
LlaMA	Seen	7B	0.51	9.3	2.01	8.0	77.8	15.7	23.1	11.0	0.17	0.13	0.12	0.10	877
LlaMA2	Seen	7B	1.87	24.3	6.03	15.1	82.5	19.0	28.1	21.4	1.10	0.92	0.85	0.84	357
Mistral	Seen	7B	1.91	25.1	6.40	15.2	82.6	20.3	29.5	22.5	1.33	1.11	1.03	0.98	334
Vicuna-7B	Seen	7B	0.98	19.6	4.42	12.3	82.6	10.0	15.9	11.8	1.03	0.92	0.86	0.83	111
Vicuna-13B	Seen	13B	1.91	25.1	6.37	15.2	82.6	20.3	29.5	22.5	1.33	1.11	1.03	0.98	334
Qwen-Chat	Seen	7B	0.62	11.9	2.47	7.3	79.4	7.4	11.7	8.3	0.64	0.52	0.51	0.48	104
Yi-6B-Chat	Seen	6B	0.99	14.6	3.74	11.2	79.6	13.9	21.3	12.6	0.64	0.60	0.57	0.55	698
Yi-34B-Chat	Seen	34B	1.00	12.9	3.41	10.3	79.0	17.6	24.8	12.7	0.92	0.85	0.81	0.79	750
GPT-4	Seen	-	2.20	26.0	6.90	14.8	82.5	29.7	38.3	31.0	2.50	2.30	2.32	2.31	369

Table 10: Comprehensive Performance of Base Language Models with Title Integration. This table showcases the performance of primary models, both featured and not featured in the main analysis, across 'seen' and 'unseen' settings, evaluated using additional metrics such as BLEU, BERTscore, and ROUGE.

	mPlug_owl2	LlaVA-NeXT (Vicuna13B)	LlaVA-NeXT (Vicuna7B)	LLaVA-NeXT (Yi34B)	Qwen-VL-Chat	Qwen-VL-Chat (FT)	GPT-4-Vision
Exact match	1.6%	0.0%	0.0%	0.0%	4.0%	5.7%	**8.97%**
Partial match	54.2%	39.9%	27.5%	66.3%	53.6%	**66.7%**	64.0%

Table 11: LVLM Primary Group Analysis of Title Generation Accuracy from Image Information.

Setting	BLIP2 (OPT)	BLIP2 (FLAN-T5-XL)	BLIP2 (FLAN-T5-XXL)	mPLUG_Owl	LLaVA-1.5	InstructBLIP (FLAN-T5-XL)
Exact match	0.0%	1.04%	1.25%	1.97%	0.0%	0.93%
Partial match	0.10%	49.6%	49.1%	37.0%	40.3%	44.0%

Table 12: LVLM Complementary Group Analysis of Title Generation Accuracy Using Only Image Information (Part 1).

Setting	InstructBLIP (FLAN-T5-XXL)	InstructBLIP (Vicuna-7B)	Instruct Blip (Vicuna-13B)	LLaVA-NeXT (mistral)	Yi-VL-6B
Exact match	1.04%	1.14%	1.14%	0.10%	1.36%
Partial match	50.1%	50.5%	58.1%	47.7%	50.6%

Table 13: LVLM Complementary Group Analysis of Title Generation Accuracy Using Only Image Information (Part 2).

Title	Rank	mPLUG-Owl	mPLUG-Owl2	Qwen-VL-Chat	Qwen-VL-Chat(FT)	GPT-4-Vision
Mona Lisa	1	✔	✔	✔	✔	✔
The Great Wave off Kanagawa	2	✔	✔		✔	✔
Vitruvian Man	3	✔	✔	✔	✔	✔
Winged Victory of Samothrace	4	✔			✔	✔
Girl with a Pearl Earring	5	✔	✔	✔	✔	✔
The Wedding at Cana	6	✔		✔	✔	✔
The Anatomy Lesson of Dr. Nicolaes Tulp	7	✔			✔	✔
Apollo Belvedere	9	✔	✔	✔		
Homeless Jesus	11			✔	✔	✔
Raphael Rooms	12					✔
Almond Blossoms	13	✔	✔			✔
The Death of General Wolfe	14	✔		✔	✔	✔
The Persistence of Memory	15	✔	✔	✔	✔	✔
Doni Tondo	19					✔
The Turkish Bath	20			✔		✔
Look Mickey	26	✔	✔	✔	✔	✔
The Seven Deadly Sins and the Four Last Things	27	✔		✔	✔	✔
The Conspiracy of Claudius Civilis	28					✔
La Belle Ferronnière	31					✔
The Gross Clinic	32				✔	✔
The Wedding Dance	33			✔	✔	✔
Sacred and Profane Love	35					✔
The Sea of Ice	37			✔	✔	
The Geographer	41			✔		✔
Equestrian Portrait of Charles V	45				✔	
The Monk by the Sea	49			✔		
My Bed	51			✔	✔	✔
I Saw the Figure 5 in Gold	55					✔
Peace Monument	57					✔
Littlefield Fountain	58				✔	✔
Music in the Tuileries	59					✔
The Cornfield	60				✔	✔
Lovejoy Columns	62			✔	✔	✔
The Allegory of Good and Bad Government	64					✔
Sibelius Monument	72			✔	✔	✔
Headington Shark	73					✔
The Great Masturbator	75					✔
Self-Portrait with Thorn Necklace and Hummingbird	81				✔	
Snow Storm: Steam-Boat off a Harbour's Mouth	83					✔
Bathers at Asnières	84				✔	✔
The Bacchanal of the Andrians	91			✔	✔	
The Painter's Studio	95				✔	
Carnation, Lily, Lily, Rose	97			✔		✔
Lady Writing a Letter with her Maid	99				✔	✔
Two Sisters (On the Terrace)	104			✔	✔	✔
Lion of Belfort	112					✔
Metamorphosis of Narcissus	114					✔
Lady Seated at a Virginal	115				✔	
Puerta de Alcalá	116				✔	✔
The Three Crosses	118			✔		
Statue of Paddington Bear	119				✔	
Our English Coasts	139					✔
Hahn/Cock	140					✔
The Wounded Deer	144			✔		✔
The Disrobing of Christ	148			✔	✔	
Lion of Venice	149			✔	✔	✔
Cross in the Mountains	153					✔
Man Writing a Letter	164		✔	✔		
Dying Slave	165					✔
Nymphs and Satyr	168	✔				
Tomb of Pope Alexander VII	172				✔	
Greece on the Ruins of Missolonghi	178					✔
The Basket of Apples	186				✔	✔
James Scott Memorial Fountain	189					✔
The Death of General Mercer at the Battle of Princeton, January 3, 1777	193					✔
Madonna of the Rabbit	200				✔	✔
Pyramid of Skulls	209					✔
Ascending and Descending	220					✔
The Madonna of Port Lligat	221				✔	✔
Le Pont de l'Europe	231					✔

Continued on next page

Table 14 – continued from previous page

Title	Rank	mPLUG-Owl	mPLUG-Owl2	Qwen-VL-Chat	Qwen-VL-Chat(FT)	GPT-4-Vision
Bratatat!	240				✔	
Marie Antoinette with a Rose	247			✔	✔	✔
The Beguiling of Merlin	256			✔	✔	
Blob Tree	258	✔	✔	✔	✔	✔
Morning in a Pine Forest	266				✔	✔
Swann Memorial Fountain	271					✔
Equestrian Portrait of Philip IV	272				✔	
Golden Guitar	274		✔	✔	✔	✔
The Blind Girl	275					✔
The Lament for Icarus	278					✔
Love's Messenger	289					✔
Arrangement in Grey and Black, No. 2: Portrait of Thomas Carlyle	304			✔		
The Return of the Herd	320					✔
Statue of Henry W. Grady	327					✔
Young Ladies of the Village	333					✔
Why Born Enslaved!	355					✔
Apollo Pavilion	358					✔
Looking Into My Dreams, Awilda	371					✔
Australian Farmer	378	✔	✔	✔	✔	✔
Bust of Giuseppe Mazzini	379					✔
Wind from the Sea	399			✔	✔	
Art is a Business	415	✔	✔			
Statue of George M. Cohan	417	✔		✔		
The Union of Earth and Water	434					✔
Frederick the Great Playing the Flute at Sanssouci	440					✔
Procession in St. Mark's Square	441					✔
Larry La Trobe	443					✔
From this moment despair ends and tactics begin	460			✔	✔	
Winter Landscape with Skaters	479				✔	
Bust of William H. English	489		✔			✔
Statue of Roscoe Conkling	507					✔
Still Life and Street	531					✔
Statue of William Blackstone	536			✔		
Statue of Chick Hearn	558				✔	
Happy Rock	587	✔	✔	✔	✔	✔
The Revells of Christendome	608				✔	
Bust of Cardinal Richelieu	629					✔
Stag Hunt	634			✔		
The Drover's Wife	679				✔	
My Egypt	684					✔
The Viaduct at L'Estaque	731					✔
The Repast of the Lion	733					✔
Puget Sound on the Pacific Coast	761					✔
Diana and Cupid	768				✔	✔
Portrait of Cardinal Richelieu	778				✔	
Statue of Toribio Losoya	873				✔	
Statue of Valentín Gómez Farías	877				✔	

Table 14: List of titles that were actually output by the model with exact settings.

Type	Template
Template 1	
Section	Focus on **{title}** and explore the {section}.
Subsection	In the context of **{title}**, explore the {subsection} of the {section}.
Sub subsection	Focusing on the {section} of **{title}**, explore the {subsubsection} about the {subsection}.
Template 2	
Section	Focus on **{title}** and explain the {section}.
Subsection	In the context of **{title}**, explain the {subsection} of the {section}.
Sub subsection	Focusing on the {section} of **{title}**, explain the {subsubsection} about the {subsection}.
Template 3	
Section	Explore the {section} of this artwork, **{title}**.
Subsection	Explore the {subsection} about the {section} of this artwork, **{title}**.
Sub subsection	Explore the {subsubsection} about the {subsection} of the {section} in this artwork, **{title}**.
Template 4	
Section	Focus on **{title}** and discuss the {section}.
Subsection	In the context of **{title}**, discuss the {subsection} of the {section}.
Sub subsection	Focusing on the {section} of **{title}**, discuss the {subsubsection} about the {subsection}.
Template 5	
Section	How does **{title}** elucidate its {section}?
Subsection	In **{title}**, how is the {subsection} of the {section} elucidated?
Sub subsection	Regarding **{title}**, how does the {section}'s {subsection} incorporate the {subsubsection}?
Template 6	
Section	Focus on **{title}** and analyze the {section}.
Subsection	In the context of **{title}**, analyze the {subsection} of the {section}.
Sub subsection	Focusing on the {section} of **{title}**, analyze the {subsubsection} about the {subsection}.
Template 7	
Section	In **{title}**, how is the {section} discussed?
Subsection	Describe the characteristics of the {subsection} in **{title}**'s {section}.
Sub subsection	When looking at the {section} of **{title}**, how do you discuss its {subsection}'s {subsubsection}?

Table 15: Prompt Templates.

```
{
  "id": "0001_T",
  "title": "Mona Lisa",
  "conversations": [
    {
      "from": "user",
      "value": "<img>/images/Mona Lisa.jpg</img>\nFocus on Mona Lisa and explore the
     history."
    },
    {
      "from": "assistant",
      "value": "Of Leonardo da  V i n c i s  works, the Mona Lisa is the only portrait
    whose authenticity...."
    }
  ]
}
```

Figure 5: Train set format with title.

```
{
  "id": "0001_NT",
  "conversations": [
    {
      "from": "user",
      "value": "<img>/images/Mona Lisa.jpg</img>\nFocus on this artwork and explore the history."
    },
    {
      "from": "assistant",
      "value": "Of Leonardo da V i n c i s works, the Mona Lisa is the only portrait whose authenticity...."
    }
  ]
}
```

Figure 6: Train set format without title.

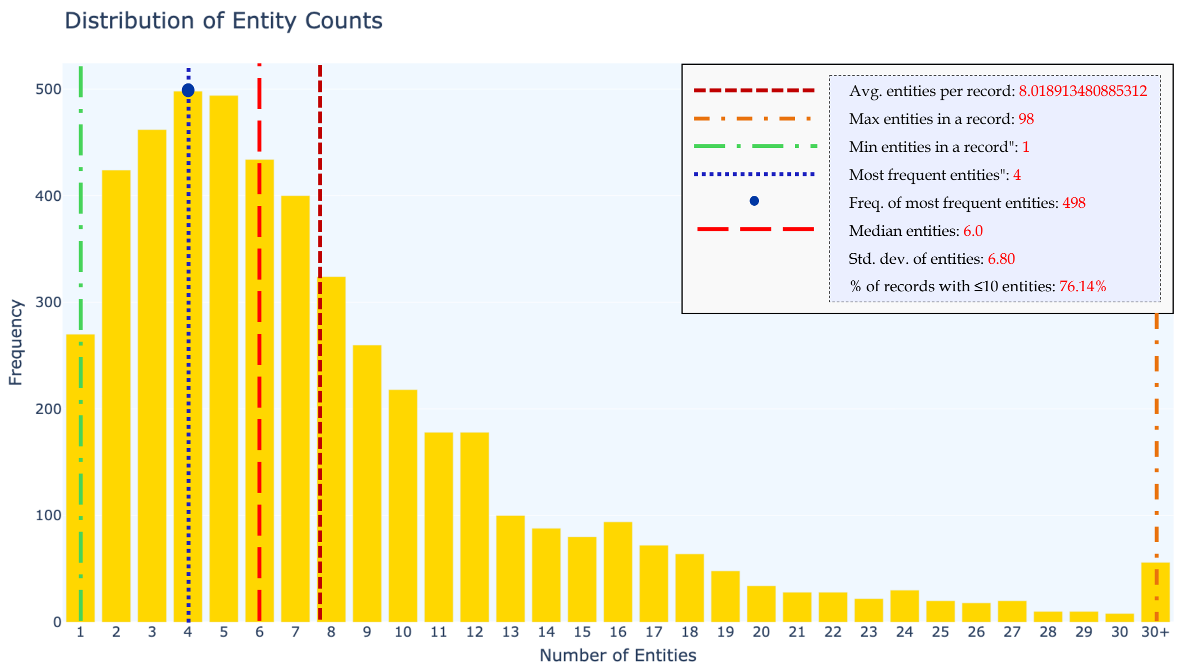

Figure 7: Entity distribution within each dataset under the ’with title’ setting.

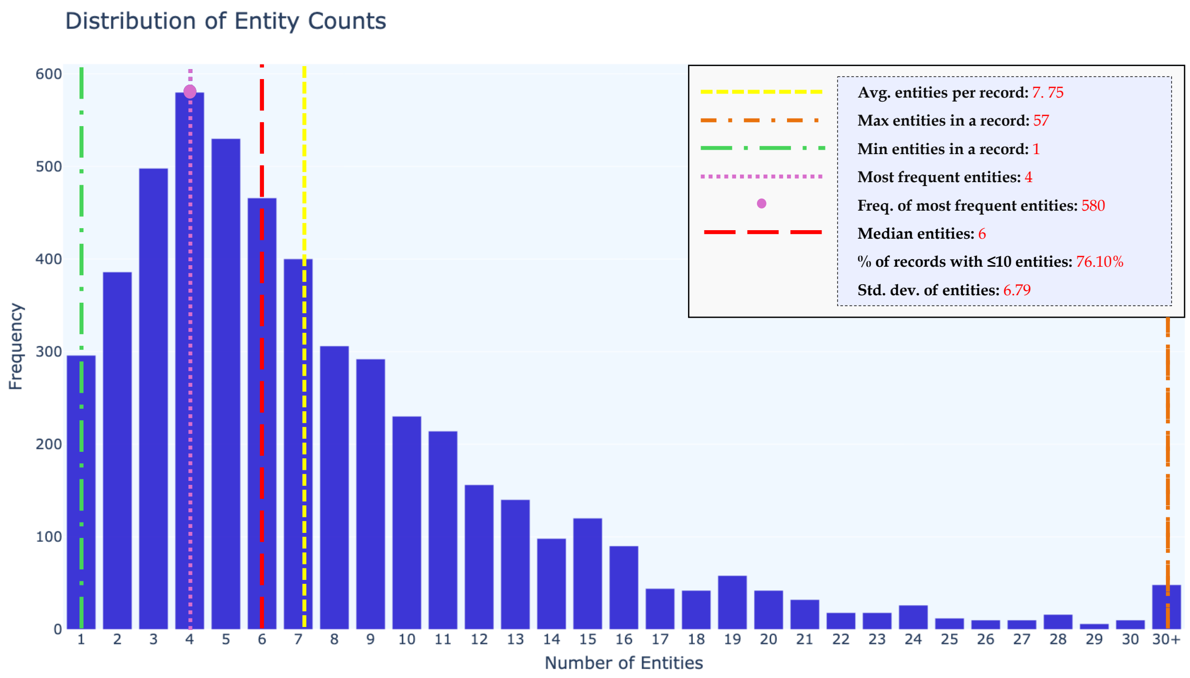

Figure 8: Entity distribution within each dataset under the 'without title' setting.

Data Type	Data Name	mPlug-owl	Qwen-VL-Chat	LLava-v-1.5	InstructBLIP
Text	ShareGPT (Chen et al., 2023)	✔		✔	
	SlimOrca (Mukherjee et al., 2023)	✔			
	In-house Data		✔		
Dialogue	LLaVA (Liu et al., 2023b)	✔		✔	
Caption	COCO (Lin et al., 2014)	✔	✔		✔
	TextCaps (Sidorov et al., 2020)	✔		✔	✔
	SBU (Yago et al., 2016)		✔		
	Coyo (Byeon et al., 2022)		✔		
	DataComp (Samir Yitzhak Gadre, 2023)		✔		
	CC12M & 3M (Changpinyo et al., 2021)		✔		
	LAION-en (Schuhmann et al., 2022) & zh		✔		
VQA	VQAv2	✔	✔	✔	✔
	GQA (Hudson and Manning, 2019)	✔	✔	✔	✔
	OKVQA (Marino et al., 2019)	✔		✔	✔
	OCRVQA (Mishra et al., 2019)	✔	✔	✔	✔
	A-OKVQA (Schwenk et al., 2022)	✔		✔	✔
	DVQA (Kafle et al., 2018)		✔		
	TextVQA (Singh et al., 2019)		✔	✔	✔
	ChartQA (Masry et al., 2022)		✔		
	A12D		✔		
Grounding[2]	GRIT (Peng et al., 2023)		✔		
Ref Grounding	GRIT		✔		
	VisualGenome (Krishna et al., 2017)		✔	✔	
	RefCOCO (Yu et al., 2016)		✔	✔	
	RefCOCO+ (Yu et al., 2016)		✔	✔	
	RefCOCOg		✔	✔	
OCR	SynthDoG-en (Kim et al., 2022) & zh		✔		
	Common Crawl pdf & HTML		✔		
Image Captioning	Web CapFilt (Li et al., 2022b)				✔
	NoCaps				✔
	Flickr30K (Hambardzumyan et al., 2023)				✔
Visual Spatial Reasoning	IconQA (Lu et al., 2021)				✔
Visual Dialog	Visual Dialog				✔
Video Question Answering	MSVD-QA (Xu et al.)				✔
	MSRVTT-QA				✔
	iVQA (Liu et al., 2018)				✔
Image Classification	VizWiz (Gurari et al., 2018)				✔
Knowledge-Grounded Image QA	ScienceQA (Lu et al., 2022)				✔

Data Type	Data Name	mPLUG-Owl2	Qwen-VL-Chat	LLava-v-1.5	InstructBLIP

Table 16: Details of training datasets.