Towards Cross-Lingual Explanation of Artwork in Large-scale Vision Language Models

Shintaro Ozaki[α] **Kazuki Hayashi**[α] **Yusuke Sakai**[α]
Hidetaka Kamigaito[α] **Katsuhiko Hayashi**[β] **Taro Watanabe**[α]

[α] Nara Institute of Science and Technology (NAIST) [β] The University of Tokyo
{ozaki.shintaro.ou6, kamigaito.h, taro.watanabe}@naist.ac.jp

Abstract

As the performance of Large-scale Vision Language Models (LVLMs) improves, they are increasingly capable of responding in multiple languages, and there is an expectation that the demand for explanations generated by LVLMs will grow. However, pre-training of Vision Encoder and the integrated training of LLMs with Vision Encoder are mainly conducted using English training data, leaving it uncertain whether LVLMs can completely handle their potential when generating explanations in languages other than English. In addition, multilingual QA benchmarks that create datasets using machine translation have cultural differences and biases, remaining issues for use as evaluation tasks. To address these challenges, this study created an extended dataset in multiple languages without relying on machine translation. This dataset that takes into account nuances and country-specific phrases was then used to evaluate the generation explanation abilities of LVLMs. Furthermore, this study examined whether Instruction-Tuning in resource-rich English improves performance in other languages. Our findings indicate that LVLMs perform worse in languages other than English compared to English. In addition, it was observed that LVLMs struggle to effectively manage the knowledge learned from English data[1].

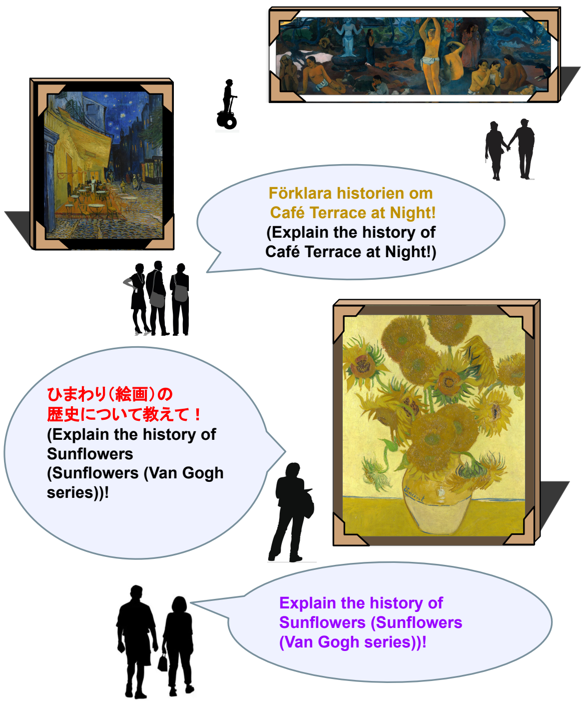

Figure 1: An example of situations that require multilingual and explanation skills.

1 Introduction

Each artwork, e.g., image, has a unique title, making it suitable for evaluating Large-scale Vision Language Models (LVLMs) that handle both the image and the text. Hayashi et al. (2024) focused on artwork explanation generation to investigate the relationship between language-based and vision-based knowledge of LVLMs using English data. When using LVLMs for creative support, explanation generation abilities are required based on the composition and ingenuity of the image, e.g., comparisons with other works, historical background, and deep artistic knowledge. LVLMs enable image and text aware tasks exactly, e.g., determining the color of traffic lights in the image and judging if it is possible to proceed, by integrating Vision Encoder (Junnan et al., 2023), e.g., Vision Transformer (ViT) (Alexey et al., 2020), which processes image data into high-dimensional features, and Large Language Models (LLMs) (LLM-jp et al., 2024), which can handle natural language, through additional training. This allows LVLMs to understand instructions with image inputs by humans and generate responses based on those instructions and they have archived remarkable performance on Vision & Language (V&L) benchmarks (Yuan et al., 2023; Bohao et al., 2023; Ozaki et al., 2024a).

[1] Our data is publicly available at https://huggingface.co/datasets/naist-nlp/MultiExpArt.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 3773–3809
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

However, there are remaining issues with training current LVLMs when dealing with multilingual data. Training and evaluation of LVLMs often use English data, leaving questions on the performance on other languages, and there exists no standard protocol especially when evaluating the performance of multilingual image understanding tasks. Several multilingual image QA tasks do exist (Soravit et al., 2023; Luu-Thuy et al., 2023), but they primarily rely on machine translation, making it uncertain whether country-specific cultural nuances or biases are completely considered. Specifically, when creating multilingual QA tasks, Yusuke et al. (2024) pointed out that multiple concepts e.g., 'roast', 'grill', 'broil', 'toast', and 'bake' in English could be potentially translated into only one expression e.g., '焼く' in Japanese. Thus, it is necessary to construct a completely fair multilingual evaluation dataset for explanation generation abilities. The issue is, in particular, compounded in the field of art, since an explanation of an image may vary across countries, leading to different explanations due to the impression of the image in other countries. Simply translating from resource-rich languages like English into other languages using machine translation to create datasets fails to account for cultural nuances. For example, "Mona Lisa" is translated directly into Chinese and Japanese correctly, but in Spanish, it is translated as "Mona Lisa" even though it is called "La Gioconda" in Spanish. Moreover, since these QA datasets do not evaluate the ability to generate explanations, there are no appropriate metrics to evaluate the explanation generation abilities of LVLMs across different languages.

To solve the lack of datasets that can evaluate the ability to generate explanations in other languages and the inability to account for country-specific cultural nuances or biases by simply using machine translation to create datasets, we created datasets that allow you to evaluate the ability to generate explanations in other languages without machine translation using Wikipedia. Hayashi et al. (2024) focused only on English, but our study expanded this work to ten languages (Chinese, Dutch, English, French, German, Italian, Japanese, Russian, Spanish, and Swedish).

We utilized these datasets to analyze the multilingual performance of current LVLMs in generation explanation abilities related to artworks with three settings which are Alignment-10, Alignment-5, and Full tasks, and investigated whether LVLMs can maintain equal generation explanation abilities in artworks when extended to ten languages. To investigate LVLMs' multilingual generation explanation abilities, we hypothesized that "the integrated training of LVLMs and the pre-training of Vision Encoder are mainly trained in English data, limiting their ability to achieve optimal performance when handling other languages." Moreover, we also conducted Instruction-Tuning in English-only training data for two of the models so that validate the extent to which these two models can acquire explanation generation capabilities in other languages solely from English training data.

We found that LVLMs perform best when given instructions in English and generating output in English, while their performance declines when instructions or output are in languages other than English. Moreover, we observed that outputting in the same language as the instructions like Japanese instruction with its Japanese response leads to better performance than the response in English for Japanese instruction, indicating that LVLMs struggle to effectively utilize the knowledge learned in English when applied to other languages. The result also showed that performance was further worse with Instruction-Tuning conducted in English. These findings support our hypothesis and suggests that it is necessary to let Vision Encoder train not only English training data but also other language data.

2 Related Work

LVLMs In general, an LVLM comprises a Vision Encoder that processes visual information and an LLM pre-trained on a large amount of textual data. They are trained using contrastive learning (Ting et al., 2020), aiming to integrate visual and linguistic information. Vision Encoder is a model trained to encode images and visual data, typically using architectures such as ResNet (Kaiming et al., 2015) or Vision Transformer (ViT) (Alexey et al., 2020). On the other hand, LLMs are models pre-trained on a large text dataset, with prominent examples including Qwen (Bai et al., 2023; Yang et al., 2024), LLaMA (Hugo et al., 2023b,a; Abhimanyu et al., 2024), Gemini (Gemini et al., 2023; Machel et al., 2024) and GPT (Tom et al., 2020; Long et al., 2022; Tim et al., 2022) LVLMs such as Qwen-VL (Jinze et al., 2023a), and LLaVA-NeXT (Haotian et al., 2024) are examples of integrated models. These

Language	Type	Template	Instruction	Output
English	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...
Japanese	Section	{Title}の作品に関して、この作品の{Section}を説明してください。	モナリザの作品に関して、この作品の歴史について説明してください。	レオナルド・ダ・ヴィンチの作品の中で、「モナ・リザ」は唯一、その真偽が不確かな肖像画であり...
	Subsection	{Title}の作品に関して、この作品の{Section}に関する{Subsection}を説明してください。	モナリザの作品に関して、この作品の歴史に関する制作と日付を説明してください。	1517年10月のルイ・ド・アラゴンの訪問の記録には、「モナ・リザ」について...
Chinese	Section	解释这件艺术品的{Section}，{Title}。	解释这件艺术品的历史，蒙娜丽莎。	在达芬奇的作品中，蒙娜丽莎是唯一一幅真伪有争议的肖像画...
	Subsection	解释关于这件艺术品的{Section}的{Subsection}，{Title}。	解释关于这件艺术品的历史的创作和日期，蒙娜丽莎。	路易·德·阿拉贡在1517年10月访问的记录中提到，"蒙娜丽莎"...
Spanish	Section	Explica la {Seccion} de esta obra de arte, {Title}.	Explica la Historia de esta obra de arte, Mona Lisa.	De las obras de Leonardo da Vinci, la Mona Lisa es el único retrato cuya autenticidad...
	Subsection	Explica la {Subsection} sobre la {Section} de esta obra de arte, {Title}.	Explica la Creación y fecha sobre la Historia de esta obra de arte, Mona Lisa.	El registro de una visita en octubre de 1517 de Luis de Aragón menciona que la Mona Lisa...

Table 1: Examples of templates and instructions for the proposed task. The blue part indicates the artwork's title and the red part indicates the names of sections and subsections in the original Wikipedia articles that correspond to their explanations. We prepared such templates for ten languages and asked native speakers to make sure they are on the same level as English.

models achieve visual and natural language integration by acquiring features from images through the Vision Encoder and textual features through LLMs and then performing additional training with the goal of integrating vision and language.

LVLMs & Knowledge Whether the visual knowledge learned by the Vision Encoder and the linguistic knowledge learned by LLMs are properly aligned remains mostly unclear (Junnan et al., 2022, 2023). Especially for generating explanations involving knowledge about artwork, which this study focuses on, it is essential to systematically align and utilize both types of knowledge (Hayashi et al., 2024). This requires the integration of visual knowledge (e.g., visual features of specific artworks) and linguistic knowledge (e.g., historical background and technical details about those artworks). In LVLMs, the integration of Vision Encoder and LLMs are achieved by adding partial networks, but this alone makes it challenging to properly align visual and linguistic knowledge. In domains requiring sophisticated knowledge, such as artwork, improper alignment can degrade the quality of generated explanations. Thus, while this study aims to integrate visual and linguistic information and build efficient models using contrastive learning, it also indicates that further research is necessary to achieve proper alignment of visual and linguistic knowledge.

LVLMs & Multilingual As we mentioned earlier, LVLMs follow human instructions through integrated learning of Vision Encoder (Junnan et al., 2023) and LLMs trained by a large amount of English training data. However, it is unclear whether LVLMs are able to really understand and output properly when input from languages other than English. On the other hand, as far as evaluation tasks such as XGQA (Jonas et al., 2022) they expanded the English GQA dataset into seven languages through translation. However, because this expansion relies on translations from English, it likely includes QA pairs that do not consider the cultural contexts of the target languages. For instance, MaXM (Soravit et al., 2023) collects large data sets by translating non-English language data into English, which is then back-translated into seven languages. Similarly, EVJVQA (Luu-Thuy et al., 2023) creates around 33,000 QA pairs from approximately 5,000 images taken in Vietnam, but the translations still retain biases unique to Viet-

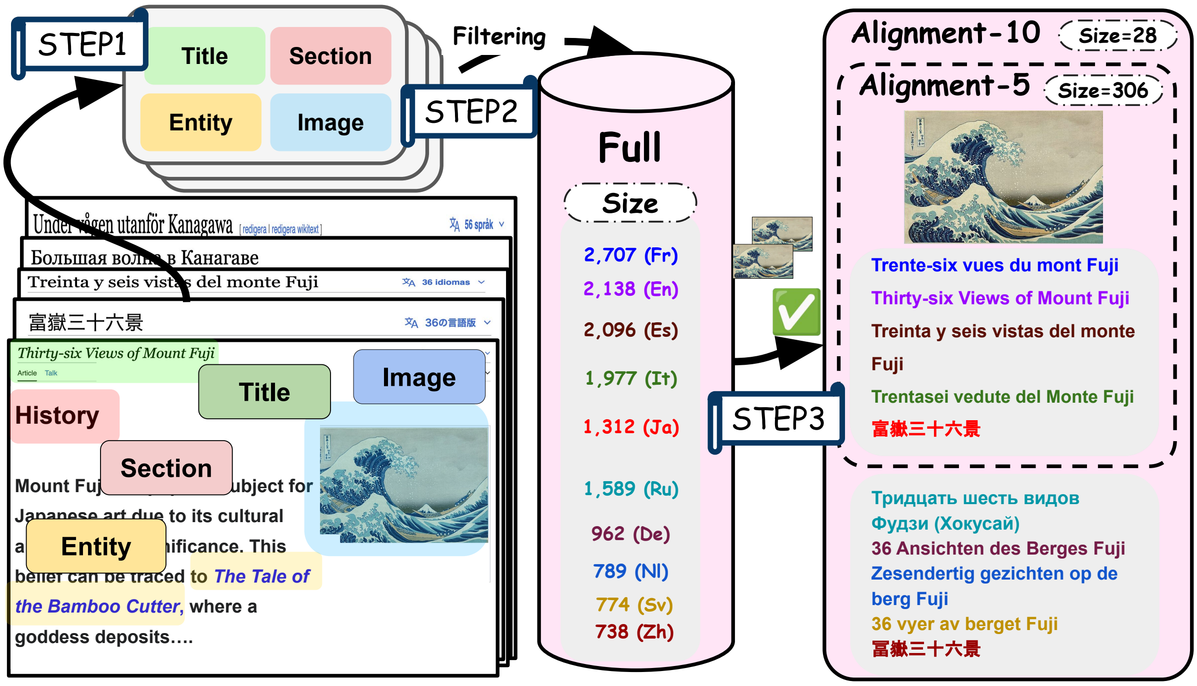

Figure 2: How to make datasets from Wikipedia. As shown in Section 4, we extracted and filtered Wikipedia pages about artworks. We then manually identified pages with titles and images common across ten languages.

namese culture and norms. In our research, we mitigated these biases by focusing on artworks, preventing the introduction of a specific culture to any country within the images. (i.e., There are countries where cars drive on the right lane and others where they drive on the left.) Since artworks have unique and definitive relationships between the title and its image, we also create datasets from relatively resource-rich Wikipedia in various languages without relying on machine translation. Our study is not a Question Answering task, such as VQA (Antol et al., 2015), but an explanation task, which requires LVLMs to explain images correctly. We evaluated an explanation-generaton task in ten languages expanding Hayashi et al. (2024).

3 Task

Our task is "Analyzing the multilingual performance of LVLMs in explaining artworks". To tackle this, we measured explanatory capabilities using three settings (Alignment-10, Alignment-5, Full) which were described below. In addition, we prepared templates for input when evaluating explanation abilities with datasets we created.

Alignment-10 We created datasets composed only of data with the same images and titles across ten languages from the Full task, which have a total of 28 pages. In other words, this dataset contains the same images with titles represented in the language of each country, allowing for an equal evaluation of description generation capabilities across the ten languages.

Alignment-5 To mitigate the data scarcity issue in Alignment-10, Alignment-5 restricts the target languages to five specific languages. The total number of data is 306, and this dataset is used to compare explanation generation abilities across the five languages. To cover a diverse range of language families, we selected English, Spanish, French, Italian, and Japanese as Alignment-5 task.

Full To further mitigate the data scarcity issues in the above settings, Full ignores the correspondence of artworks between languages and treats each language independently. For details on the number of data, refer to Table 7 or Figure 4. By using the Full task, we aimed to evaluate the differences in performance.

Templates We prepared templates for evaluating explanation generation abilities using the datasets created from three tasks mentioned above. The process is as follows: 1) We prepared four patterns of templates for each of ten languages to mitigate the variance issue in performance evaluation (Sakai et al., 2024b). In templates, we referred to the study by Hayashi et al. (2024), selecting four pat-

Input	Output	LVLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
			exact	partial		n=0	n=1	n=2	n=∞	
En	En	LLaVA-NeXT	**26.49**	**31.54**	**26.07**	1.35	1.65	1.66	1.70	252
		LLaVA-NeXT (FT)	16.98	22.70	19.95	3.02	3.33	3.23	3.11	83
		Qwen-VL	21.11	27.03	21.78	1.60	1.59	1.56	1.52	155
		Qwen-VL (FT)	21.12	24.87	21.95	**3.57**	**3.83**	**3.78**	**3.68**	177
		mPLUG-Owl2	12.79	17.08	13.48	2.07	1.68	1.59	1.56	151

Table 2: Results of LVLMs in Alignment-10 Task (the instruction and the output in English, {En}-{En}). Bold fonts indicate the best scores. The red and blue figures shown in the following figures are the different figures compared to this Table. "(FT)" indicates the model conducted LoRA-Tuning.

terns with clearly different grammatical structures to avoid a lack of diversity. Yusuke et al. (2024) noted that not choosing distinctly different patterns may result in differences originally present in English being lost in translation; 2) We let ChatGPT[2] translate the obtained templates into ten languages. We chose to use LLMs rather than translation tools because LLMs are thought to better understand and translate including nuances; 3) Even with translations taking into nuances by ChatGPT, there may be variations in quality between languages. To solve this, we asked nine native speakers of ten languages, to check whether the templates translated back into English maintained the same nuance and level of difficulty. This process ensured that all 10 language templates created in this study have the same level of difficulty; Of course, it might be possible to crowdsource this task using platforms like MTurk[3], but asking annotators simply "Is this translation correct including nuances?" may not lead to serious engagement with the translation checking task. For examples of the each language template, refer to Table 1. During inference, the model runs as many inferences as the total number of combinations of section, subsection, and subsubsection. The appendix provides the exact number of inferences in Table 7.

4 Dataset Creation

For each of ten languages, the following steps were taken to create the dataset. Ten languages were determined based on having a higher number of Wikipedia articles than the total number of articles.

STEP1: Extracting Data from Wikipedia We collected Artwork articles from the English Wikipedia Infobox. Articles with the same title in nine other languages are identified to create corresponding articles in those languages. Hyperlinked strings within the articles are extracted as entities related to artworks. The description includes four types of information: the image, the title, hierarchical information from the article (Section, Subsection, Sub subsection), and the extracted entities.

STEP2: Filtering and Formatting From the collected articles, those without images were excluded. Any articles that had domains but no actual pages on Wikipedia were also removed. This process completes the dataset used for the Full task.

STEP3: Adjusting For the Alignment-10 and Alignment-5 tasks, we extracted pages from other languages that have the same titles as the English Wikipedia pages to ensure alignment across ten or five languages. These non-English pages include the English titles in their metadata, and alignment is successful when these English titles match exactly. To eliminate differences between languages, a manual verification is conducted to ensure that all articles contain images of the same artwork. Variations in image size are permitted, but all images must represent the same artwork across languages. The datasets for Alignment-10 and Alignment-5 are prepared accordingly, using images from the English articles for alignment.

STEP4: Data Splitting To measure the explanation generation abilities of LVLMs, the following approach is used: (1) For the Alignment task, all data was treated as test set. (2) For the Full task, nine non-English languages are used for test set, while English data is divided into train, dev, and test sets. To avoid biases arising from the popularity of artworks in the LVLM's training data, we shuffled the English data based on six indicators: page views, number of links, number of edits, number of references, number of language versions, and article length (Hayashi et al., 2024). The data

[2] https://openai.com/chatgpt/
[3] https://www.mturk.com/

Input	Output	LVLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
			exact	partial		n=0	n=1	n=2	n=∞	
		LLaVA-NeXT	0.00 (-26.49)	2.24 (-29.30)	0.00 (-26.07)	0.00 (-1.35)	0.00 (-1.65)	0.00 (-1.66)	0.00 (-1.70)	137
		LLaVA-NeXT (FT)	6.23 (-10.75)	9.05 (-13.65)	6.87 (-13.08)	1.27 (-1.75)	1.21 (-2.12)	1.09 (-2.13)	1.06 (-2.05)	83
En	Es	Qwen-VL	**10.81 (-10.29)**	**15.18 (-11.85)**	**11.42 (-10.36)**	**1.48 (-0.12)**	**1.41 (-0.18)**	**1.34 (-0.22)**	**1.27 (-0.25)**	109
		Qwen-VL (FT)	4.25 (-16.87)	7.86 (-17.01)	5.40 (-16.55)	0.36 (-3.21)	0.36 (-3.47)	0.31 (-3.46)	0.29 (-3.39)	190
		mPLUG-Owl2	8.95 (-3.83)	11.95 (-5.13)	9.62 (-3.86)	0.93 (-1.14)	1.13 (-0.55)	1.07 (-0.52)	1.02 (-0.54)	108
		LLaVA-NeXT	1.00 (-25.49)	7.42 (-24.12)	1.33 (-24.74)	0.00 (-1.35)	0.00 (-1.65)	0.00 (-1.66)	0.00 (-1.70)	179
		LLaVA-NeXT (FT)	8.39 (-8.59)	11.41 (-11.29)	8.71 (-11.24)	1.43 (-1.59)	**1.99 (-1.34)**	**1.96 (-1.27)**	**1.95 (-1.16)**	68
En	Fr	Qwen-VL	**12.11 (-9.00)**	**17.23 (-9.80)**	**13.05 (-8.73)**	1.44 (-0.16)	1.45 (-0.14)	1.40 (-0.16)	1.34 (-0.18)	96
		Qwen-VL (FT)	7.19 (-13.92)	11.24 (-13.63)	8.34 (-13.61)	0.45 (-3.12)	0.90 (-2.93)	0.88 (-2.89)	0.89 (-2.79)	175
		mPLUG-Owl2	10.26 (-2.53)	15.51 (-1.57)	10.99 (-2.49)	**1.72 (-0.35)**	1.33 (-0.35)	1.20 (-0.39)	1.16 (-0.40)	109
		LLaVA-NeXT	**14.03 (-12.46)**	**17.90 (-13.64)**	**16.51 (-9.56)**	**1.73 (+0.38)**	**1.70 (+0.05)**	**1.67 (+0.01)**	**1.82 (+0.13)**	169
		LLaVA-NeXT (FT)	6.83 (-10.15)	9.54 (-13.16)	8.23 (-11.72)	0.86 (-2.15)	0.74 (-2.59)	0.77 (-2.46)	0.78 (-2.33)	82
En	De	Qwen-VL	10.64 (-10.46)	13.95 (-13.08)	13.21 (-8.56)	1.16 (-0.44)	1.24 (-0.35)	1.21 (-0.35)	1.40 (-0.12)	111
		Qwen-VL (FT)	7.98 (-13.14)	11.08 (-13.79)	9.86 (-12.09)	0.80 (-2.77)	0.65 (-3.18)	0.74 (-3.03)	0.74 (-2.94)	203
		mPLUG-Owl2	8.81 (-3.98)	12.12 (-4.97)	10.54 (-2.94)	0.72 (-1.35)	0.76 (-0.92)	0.74 (-0.85)	0.70 (-0.86)	98
		LLaVA-NeXT	**8.53 (-17.95)**	**13.33 (-18.21)**	**9.37 (-16.70)**	0.86 (-0.48)	0.87 (-0.79)	1.06 (-0.60)	1.05 (-0.65)	171
		LLaVA-NeXT (FT)	5.89 (-11.09)	8.90 (-13.80)	6.61 (-13.34)	0.96 (-2.06)	**1.32 (-2.01)**	**1.32 (-1.91)**	**1.31 (-1.80)**	66
En	It	Qwen-VL	7.23 (-13.87)	11.43 (-15.59)	8.71 (-13.06)	0.51 (-1.08)	0.62 (-0.97)	0.65 (-0.91)	0.63 (-0.89)	107
		Qwen-VL (FT)	5.51 (-15.61)	8.17 (-16.70)	6.53 (-15.42)	**1.14 (-2.44)**	0.82 (-3.01)	0.85 (-2.93)	0.84 (-2.84)	170
		mPLUG-Owl2	3.97 (-8.82)	8.50 (-8.58)	4.50 (-8.98)	0.15 (-1.92)	0.14 (-1.53)	0.16 (-1.43)	0.15 (-1.41)	107
		LLaVA-NeXT	12.21 (-14.28)	**17.83 (-13.71)**	14.60 (-11.46)	0.36 (-0.99)	1.81 (+0.15)	1.70 (+0.04)	**1.83 (+0.13)**	178
		LLaVA-NeXT (FT)	9.41 (-7.56)	15.01 (-7.69)	12.14 (-7.81)	**1.21 (-1.81)**	1.07 (-2.27)	0.91 (-2.32)	1.02 (-2.09)	119
En	Nl	Qwen-VL	11.07 (-10.04)	16.44 (-10.59)	12.73 (-9.05)	0.89 (-0.71)	**1.90 (+0.32)**	**1.78 (+0.22)**	1.80 (+0.28)	132
		Qwen-VL (FT)	**12.67 (-8.45)**	17.03 (-7.84)	**16.91 (-5.04)**	1.02 (-2.55)	0.96 (-2.88)	0.95 (-2.83)	1.01 (-2.67)	181
		mPLUG-Owl2	8.27 (-4.51)	13.46 (-3.62)	9.06 (-4.42)	0.46 (-1.61)	0.43 (-1.25)	0.41 (-1.18)	0.41 (-1.14)	100
		LLaVA-NeXT	**15.01 (-11.48)**	**18.65 (-12.89)**	**13.56 (-12.51)**	**1.29 (-0.05)**	0.97 (-0.69)	**1.15 (-0.51)**	**1.09 (-0.61)**	174
		LLaVA-NeXT (FT)	10.00 (-6.97)	12.43 (-10.27)	10.54 (-9.41)	0.84 (-2.17)	**1.08 (-2.26)**	0.97 (-2.26)	0.87 (-2.24)	115
En	Sv	Qwen-VL	10.37 (-10.74)	14.08 (-12.94)	10.15 (-11.62)	0.84 (-0.76)	0.86 (-0.72)	0.83 (-0.73)	0.80 (-0.72)	123
		Qwen-VL (FT)	8.97 (-12.14)	12.25 (-12.61)	9.66 (-12.29)	0.87 (-2.70)	0.94 (-2.89)	0.92 (-2.86)	0.90 (-2.78)	164
		mPLUG-Owl2	10.21 (-2.57)	13.03 (-4.05)	9.07 (-4.41)	0.35 (-1.72)	0.35 (-1.33)	0.34 (-1.25)	0.34 (-1.22)	88
		LLaVA-NeXT	**10.32 (-16.17)**	**15.15 (-16.39)**	**8.53 (-17.54)**	**0.32 (-1.02)**	**0.36 (-1.30)**	**0.31 (-1.35)**	**0.32 (-1.38)**	203
		LLaVA-NeXT (FT)	0.55 (-16.42)	1.87 (-20.83)	0.49 (-19.46)	0.00 (-3.02)	0.02 (-3.32)	0.02 (-3.21)	0.01 (-3.10)	85
En	Ru	Qwen-VL	4.59 (-16.52)	8.05 (-18.97)	3.51 (-18.26)	0.02 (-1.58)	0.07 (-1.52)	0.07 (-1.49)	0.07 (-1.45)	113
		Qwen-VL (FT)	0.00 (-21.12)	0.95 (-23.91)	0.00 (-21.95)	0.00 (-3.57)	0.00 (-3.83)	0.00 (-3.78)	0.00 (-3.68)	169
		mPLUG-Owl2	5.99 (-6.80)	8.68 (-8.40)	4.88 (-8.60)	0.00 (-2.07)	0.02 (-1.66)	0.01 (-1.57)	0.01 (-1.54)	99
		LLaVA-NeXT	**8.68 (-17.81)**	**8.68 (-22.86)**	**11.47 (-14.60)**	**0.80 (-0.54)**	**0.80 (-0.85)**	**0.80 (-0.86)**	**0.80 (-0.90)**	211
		LLaVA-NeXT (FT)	0.29 (-16.68)	0.30 (-22.40)	0.38 (-19.57)	0.04 (-2.98)	0.04 (-3.29)	0.04 (-3.19)	0.04 (-3.07)	85
En	Ja	Qwen-VL	3.52 (-17.59)	3.53 (-23.49)	4.78 (-17.00)	0.32 (-1.28)	0.32 (-1.27)	0.32 (-1.24)	0.32 (-1.20)	132
		Qwen-VL (FT)	0.00 (-21.12)	0.03 (-24.84)	0.00 (-21.95)	0.00 (-3.57)	0.00 (-3.83)	0.00 (-3.78)	0.00 (-3.68)	188
		mPLUG-Owl2	3.75 (-9.04)	3.75 (-13.33)	4.98 (-8.49)	0.39 (-1.68)	0.39 (-1.28)	0.39 (-1.20)	0.39 (-1.17)	112
		LLaVA-NeXT	**14.00 (-1.86)**	**14.09 (-6.86)**	**16.69 (+0.19)**	0.66 (-0.42)	0.66 (-0.58)	0.66 (-0.56)	0.66 (-0.59)	228
		LLaVA-NeXT (FT)	0.14 (-11.49)	0.39 (-15.08)	0.15 (-13.97)	0.00 (-2.42)	0.00 (-2.60)	0.00 (-2.51)	0.00 (-2.43)	92
En	Zh	Qwen-VL	10.69 (-1.45)	10.70 (-5.71)	12.71 (+0.52)	**0.74 (-0.59)**	**0.73 (-0.44)**	**0.73 (-0.39)**	**0.73 (-0.35)**	138
		Qwen-VL (FT)	0.37 (-13.40)	0.75 (-16.88)	0.51 (-12.91)	0.01 (-2.96)	0.01 (-3.09)	0.01 (-3.04)	0.01 (-2.99)	154
		mPLUG-Owl2	6.38 (-6.45)	6.40 (-10.74)	7.75 (-5.77)	0.32 (-1.75)	0.32 (-1.36)	0.32 (-1.27)	0.32 (-1.24)	108

Table 3: Results of LVLMs in Alignment-10 Task ({En}-{Lang}). Bold fonts indicate the best score for that language combination. The values are noted next to the output of the difference by the same model in the method with instruction and output in English ({En}-{En}). Red indicates a higher value than that method; blue indicates a lower value.

was ranked according to these indicators, and the test, valid, and train data were split in a 2:2:6 ratio to maintain average rankings. The data used in the Alignment task was included in the test set.

5 Experiments

5.1 Evaluation Metrics

This study adopted three evaluation metrics proposed by Hayashi et al. (2024) and also described these metrics more details in Appendix E. We also utilize popular metrics in NLG for evaluation, i.e., BLEU (Kishore et al., 2002), ROUGE (Chin-Yew, 2004), and BERTScore (Tianyi et al., 2019).

Entity Coverage, Entity F1, and Entity Cooccurrence These metrics evaluate how well the generated text incorporates entities related to the artwork and how accurately it reflects the relationships between these entities proposed by Hayashi et al. (2024). Entity Coverage measures the inclusion of relevant entities in both exact and partial matches. Entity F1 assesses the frequency and appropriateness of entity usage by comparing the generated text with reference explanations, inspired by the ROUGE metric. Entity Cooccurrence goes a step further by examining how entities are contextually combined across sentences, considering their co-occurrence within the entire text, and applying brevity penalties to avoid inflated coverage in longer explanations.

5.2 Models and Others

We chose five models with relatively high performance: mPLUG-Owl2 (Qinghao et al., 2024),

Input	Output	LVLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
			exact	partial		n=0	n=1	n=2	n=∞	
Es	Es	LLaVA-NeXT	**17.26 (-9.23)**	**21.30 (-10.25)**	**17.05 (-9.01)**	2.13 (+0.78)	2.32 (+0.67)	2.17 (+0.51)	2.10 (+0.40)	186
		LLaVA-NeXT (FT)	12.82 (-4.15)	16.84 (-5.86)	12.77 (-7.18)	0.88 (-2.14)	1.03 (-2.31)	1.26 (-1.97)	1.11 (-1.99)	147
		Qwen-VL	14.68 (-6.43)	18.43 (-8.59)	14.35 (-7.43)	2.40 (+0.81)	**2.42 (+0.83)**	**2.57 (+1.01)**	**2.56 (+1.04)**	150
		Qwen-VL (FT)	4.09 (-17.02)	7.10 (-17.77)	4.52 (-17.42)	0.15 (-3.43)	0.16 (-3.68)	0.15 (-3.63)	0.14 (-3.54)	301
		mPLUG-Owl2	10.91 (-1.87)	15.06 (-2.02)	11.91 (-1.57)	**2.47 (+0.40)**	2.07 (+0.40)	2.02 (+0.44)	1.99 (+0.44)	135
Fr	Fr	LLaVA-NeXT	**24.35 (-2.14)**	**29.27 (-2.27)**	24.38 (-1.69)	0.95 (-0.40)	0.90 (-0.75)	0.88 (-0.78)	0.90 (-0.80)	211
		LLaVA-NeXT (FT)	16.63 (-0.35)	20.13 (-2.57)	16.09 (-3.86)	1.18 (-1.83)	0.93 (-2.41)	1.00 (-2.23)	0.98 (-2.13)	98
		Qwen-VL	19.38 (-1.73)	24.71 (-2.32)	18.30 (-3.47)	1.07 (-0.53)	1.03 (-0.55)	0.96 (-0.60)	0.96 (-0.56)	165
		Qwen-VL (FT)	24.15 (+3.04)	28.59 (+3.73)	**24.79 (+2.85)**	**3.83 (+0.26)**	**4.41 (+0.58)**	**4.51 (+0.73)**	**4.51 (+0.83)**	219
		mPLUG-Owl2	17.43 (+4.64)	22.48 (+5.40)	17.78 (+4.30)	0.85 (-1.22)	0.65 (-1.02)	0.75 (-0.84)	0.73 (-0.83)	158
De	De	LLaVA-NeXT	**17.45 (-9.04)**	**20.66 (-10.89)**	**21.05 (-5.02)**	2.11 (+0.77)	**2.20 (+0.55)**	**2.22 (+0.56)**	**2.11 (+0.41)**	204
		LLaVA-NeXT (FT)	10.53 (-6.44)	13.10 (-9.60)	13.32 (-6.63)	1.53 (-1.49)	1.09 (-2.25)	1.15 (-2.07)	1.16 (-1.95)	123
		Qwen-VL	15.10 (-6.00)	18.20 (-8.82)	17.97 (-3.81)	**2.12 (+0.52)**	1.99 (+0.41)	2.08 (+0.52)	1.99 (+0.47)	160
		Qwen-VL (FT)	7.74 (-13.38)	9.58 (-15.28)	9.23 (-12.72)	0.37 (-3.20)	0.40 (-3.43)	0.43 (-3.34)	0.40 (-3.28)	287
		mPLUG-Owl2	14.33 (+1.55)	17.63 (+0.55)	16.73 (+3.25)	1.99 (-0.08)	1.92 (+0.25)	1.94 (+0.35)	1.81 (+0.25)	143
It	It	LLaVA-NeXT	**10.34 (-16.14)**	**15.43 (-16.11)**	**11.33 (-14.74)**	1.16 (-0.19)	**0.93 (-0.72)**	**0.96 (-0.70)**	**0.96 (-0.74)**	185
		LLaVA-NeXT (FT)	5.73 (-11.25)	9.84 (-12.86)	6.45 (-13.50)	0.31 (-2.71)	0.25 (-3.08)	0.25 (-2.98)	0.23 (-2.88)	91
		Qwen-VL	9.97 (-11.13)	14.20 (-12.82)	11.09 (-10.68)	**1.16 (-0.44)**	0.93 (-0.65)	0.94 (-0.62)	0.90 (-0.62)	126
		Qwen-VL (FT)	3.15 (-17.96)	6.95 (-17.92)	3.42 (-18.53)	0.15 (-3.42)	0.18 (-3.65)	0.23 (-3.54)	0.21 (-3.47)	253
		mPLUG-Owl2	8.69 (-4.10)	12.66 (-4.42)	9.54 (-3.94)	0.51 (-1.56)	0.32 (-1.36)	0.35 (-1.24)	0.33 (-1.23)	111
Nl	Nl	LLaVA-NeXT	17.66 (-8.83)	23.56 (-7.99)	**19.78 (-6.28)**	0.79 (-0.56)	**3.55 (+1.89)**	3.61 (+1.95)	3.88 (+2.18)	199
		LLaVA-NeXT (FT)	15.57 (-1.40)	20.79 (-1.91)	16.87 (-3.08)	1.66 (-1.35)	3.38 (+0.05)	3.32 (+0.09)	3.47 (+0.37)	183
		Qwen-VL	**19.41 (-1.69)**	**24.45 (-2.58)**	19.65 (-2.13)	**2.13 (+0.53)**	3.27 (+1.69)	**3.89 (+2.33)**	**4.04 (+2.52)**	172
		Qwen-VL (FT)	12.68 (-8.43)	18.46 (-6.41)	16.72 (-5.22)	1.09 (-2.48)	1.66 (-2.18)	1.81 (-1.96)	1.80 (-1.88)	300
		mPLUG-Owl2	10.78 (-2.01)	15.43 (-1.66)	12.81 (-0.67)	0.15 (-1.92)	1.08 (-0.60)	1.05 (-0.54)	1.12 (-0.43)	114
Sv	Sv	LLaVA-NeXT	**27.51 (+1.02)**	**29.61 (-1.93)**	16.71 (-9.36)	2.10 (+0.75)	0.87 (-0.78)	0.89 (-0.77)	0.90 (-0.79)	206
		LLaVA-NeXT (FT)	22.83 (+5.86)	25.10 (+2.40)	12.17 (-7.78)	2.82 (-0.20)	1.11 (-2.22)	1.17 (-2.06)	1.16 (-1.94)	169
		Qwen-VL	24.02 (+2.92)	26.69 (-0.34)	**19.18 (-2.60)**	**3.60 (+2.00)**	**1.53 (-0.06)**	**1.54 (-0.02)**	**1.50 (-0.02)**	147
		Qwen-VL (FT)	16.04 (-5.07)	18.10 (-6.77)	6.15 (-15.80)	0.23 (-3.35)	0.18 (-3.65)	0.20 (-3.57)	0.21 (-3.47)	242
		mPLUG-Owl2	21.40 (+8.61)	23.51 (+6.43)	13.84 (+0.36)	2.01 (-0.06)	1.07 (-0.61)	1.06 (-0.52)	1.05 (-0.51)	111
Ru	Ru	LLaVA-NeXT	**14.38 (-12.11)**	**17.43 (-14.11)**	**9.81 (-16.26)**	0.26 (-1.08)	**0.45 (-1.20)**	**0.42 (-1.24)**	**0.41 (-1.29)**	219
		LLaVA-NeXT (FT)	10.74 (-6.24)	13.67 (-9.03)	6.55 (-13.40)	0.32 (-2.70)	0.37 (-2.96)	0.36 (-2.87)	0.36 (-2.75)	184
		Qwen-VL	6.80 (-14.31)	9.68 (-17.34)	4.63 (-17.15)	0.31 (-1.29)	0.32 (-1.27)	0.30 (-1.26)	0.31 (-1.21)	170
		Qwen-VL (FT)	1.76 (-19.35)	3.60 (-21.27)	1.52 (-20.42)	0.14 (-3.43)	0.14 (-3.69)	0.14 (-3.64)	0.14 (-3.54)	324
		mPLUG-Owl2	7.07 (-5.72)	8.92 (-8.16)	5.57 (-7.91)	**0.51 (-1.56)**	0.34 (-1.33)	0.31 (-1.28)	0.35 (-1.21)	129
Ja	Ja	LLaVA-NeXT	**13.38 (-13.11)**	**13.38 (-18.17)**	**17.68 (-8.39)**	0.73 (-0.61)	0.83 (-0.83)	0.83 (-0.83)	0.83 (-0.87)	249
		LLaVA-NeXT (FT)	7.51 (-9.46)	7.51 (-15.19)	7.80 (-12.15)	1.14 (-1.88)	1.14 (-2.19)	1.14 (-2.09)	1.14 (-1.97)	167
		Qwen-VL	10.89 (-10.22)	10.90 (-16.13)	14.56 (-7.22)	0.92 (-0.68)	0.92 (-0.67)	0.92 (-0.64)	0.92 (-0.60)	154
		Qwen-VL (FT)	0.86 (-20.26)	0.88 (-23.99)	1.12 (-20.83)	0.03 (-3.55)	0.03 (-3.81)	0.03 (-3.75)	0.03 (-3.65)	278
		mPLUG-Owl2	6.91 (-5.88)	6.93 (-10.15)	9.34 (-4.14)	**1.20 (-0.87)**	**1.21 (-0.46)**	**1.21 (-0.38)**	**1.21 (-0.35)**	144
Zh	Zh	LLaVA-NeXT	13.78 (-2.08)	13.78 (-7.17)	17.00 (+0.50)	0.54 (-0.54)	0.53 (-0.70)	0.53 (-0.69)	0.53 (-0.72)	246
		LLaVA-NeXT (FT)	6.93 (-4.71)	6.97 (-8.50)	7.31 (-6.81)	0.78 (-1.64)	0.78 (-1.83)	0.78 (-1.73)	0.78 (-1.65)	170
		Qwen-VL	**17.90 (+5.76)**	**17.90 (+1.48)**	**22.12 (+9.93)**	**3.31 (+1.97)**	**3.30 (+2.13)**	**3.30 (+2.18)**	**3.30 (+2.22)**	155
		Qwen-VL (FT)	0.22 (-13.55)	0.33 (-17.29)	0.27 (-13.16)	0.00 (-2.97)	0.00 (-3.10)	0.00 (-3.06)	0.00 (-3.00)	249
		mPLUG-Owl2	9.03 (-3.80)	9.05 (-8.08)	12.98 (-0.55)	0.77 (-1.31)	0.77 (-0.91)	0.77 (-0.82)	0.77 (-0.80)	150

Table 4: Results of LVLMs in Alignment-10 Task (the format with instruction and output in each of the ten languages, {Lang}-{Lang}). Bold fonts indicate the best score for that language combination. The values are noted next to the differences output by the same model in the format with instruction and output in English ({En}-{En}). Red indicates a higher value than {En}-{En}; blue indicates a lower value.

LLaVA-NeXT (Haotian et al., 2023a, 2024, 2023b), XComposer2 (Xiaoyi et al., 2024), Phi-3 (Marah et al., 2024), and Qwen-VL (Jinze et al., 2023a). In addition, LLaVA-NeXT and Qwen-VL were conducted LoRA Tuning (Hu et al., 2022) with English train data and included in the evaluation. Detailed experimental settings are described in Appendix A.1. This approach is based on the observation that current LLMs perform better when instructions are given in English (Afina et al., 2024a). As far as Alignment tasks, we validated four patterns of input: {En, Lang}-{En, Lang}. This indicates that when the input is English, the output can be directed to English or another language. The same thing can also be done when the input is another language, and these four patterns were tested in this study. By testing these patterns, we verify whether LVLMs perform better when supported in English, and whether having the output in English is a meaningful instruction. As far as tokenizing words, we used SpaCy[4] as a multilingual tokenizer, tokenizing each language to perform segmentation. Thus, each language is expected to be divided into optimal token units.

5.3 Results

From the experiments conducted with Alignment-10, the method let LVLMs generate in English with English ({En}-{En}) results are listed in Table 2, the method which is instruction in English and output in other languages ({En}-{Lang}) results in Table 3, and the instruction and output in other same languages ({Lang}-{Lang}) results in Table 4. The

[4] https://spacy.io/

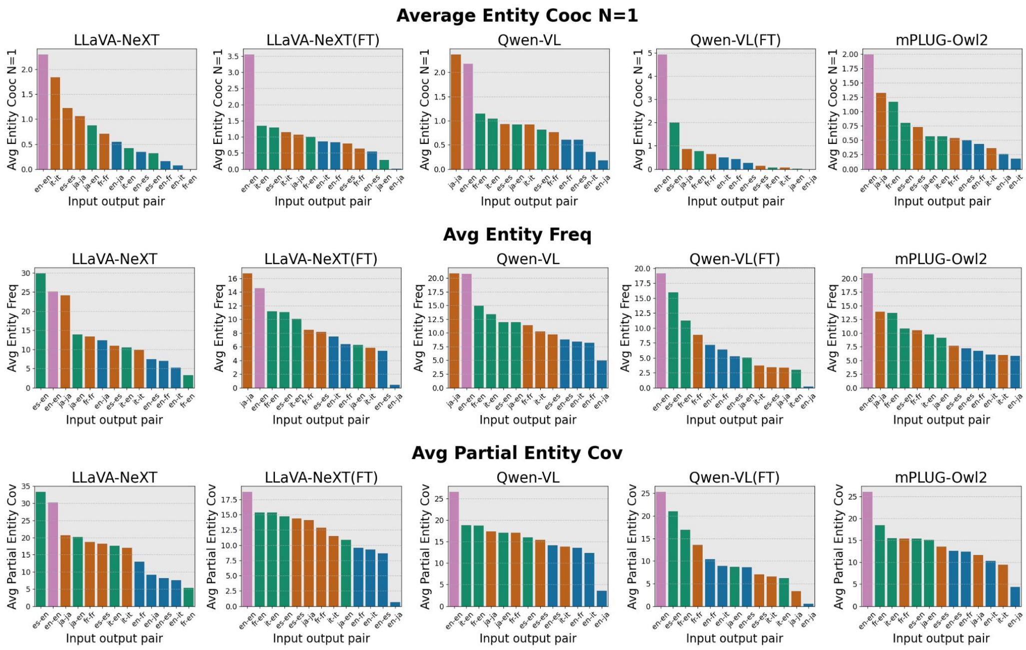

Figure 3: Some of the results in the Alignment-5 task. Purple bin indicates the method which is the instruction and the output in English ({En}-{En}), Green bin indicates the instruction in languages other than English and the output in English ({Lang}-{En}), Brown bin indicates the instruction and output in languages other than English ({Lang}-{Lang}) and Blue bin indicates the instruction in English and the output in languages other than English ({En}-{Lang}). From this figure, it can be seen that the English instructions are optimal, even if the number of data is expanded. We described further detailed results in Table 9 including Phi-3 and XComposer2. You can see the rest of the results in Figure 6 in the Appendix.

results for Phi-3 and XComposer2 are described in the Table 10 in Appendix. Overall, the results confirm that giving instructions in English and letting them generate output in English (i.e., {En}-{En}) maximizes the performance of LVLMs. On the other hand, LoRA Tuning increased the value of Entity Cooccurrance, while other values decreased. This suggests that LoRA Tuning enabled LVLMs to understand and explain the context, but prevented entities from appearing in the generated sentences. Furthermore, looking at the results of Alignment-5 in Table 9 in Appendix, where the number of data was expanded, the outputs that used English instructions and outputs were generally higher, followed by those using instructions and outputs in other languages. This is consistent with the results of Alignment-10. In addition, Figure 3 includes results where instructions were given in other languages and outputs were produced in English.

6 Analysis and Discussion

Which Instruction and Output Language is Best? We confirmed that the pattern which instruction and output are English ({En}-{En}) performed the best ability, whereas the performance is lower for the pattern in which instruction in English and output in other languages ({En}-{Lang}, i.e., Please generate the output in Chinese). This suggests that "LVLMs have a poor ability to successfully transfer knowledge learned in English to other languages". We also confirmed that this effect was more pronounced in the LoRA-Tuning model (LLaVA-NeXT(FT) and Qwen-VL(FT)).

LVLMs' Ability to Explain Artworks in Other Languages Considering the multilingual explanation generation capabilities of LVLMs, a comparison between Table 3 and Table 4 reveals that performing the method is instruction and output in other same languages ({Lang}-{Lang}) generally yields better results than in the instruction in English and output in other language ({En}-{Lang}). When explaining in the native language using data trained in that language, the model effectively manages the knowledge. However, when explaining in other languages using knowledge trained in English, the model struggles to handle the information adequately. This result shows particularly clear in the cases of Qwen-VL's results between the method is instruction and output in Chinese ({Zh}-{Zh}) and the instruction in English and output in English ({Zh}-{En}) pairs. In addition, using English training data for LoRA Tuning likely leads to the forgetting of original performance, resulting in a decline in effectiveness. From these observations, it is clear that LVLMs currently exhibit their maximum capabilities only when instructed and output in English ({En}-{En}). Thus, future research should focus on training LVLMs in multiple languages.

Comparison between En-En and Lang-Lang capabilities In this result, En-En performance was the highest. It indicates that existing LVLMs have been primarily pre-trained on English data, and providing both input and output in English allows the model to perform at its best. {Lang}-{Lang} performance was the next best. While it is evident that LVLMs have been pre-trained on languages other than English and can handle multilingual tasks, their performance still falls short of En-En. We believe this is due to insufficient pre-training on non-English languages.

Comparison between En-{Lang} and {Lang}-{Lang} We also compared En-{Lang} and {Lang}-{Lang} results, finding that {Lang}-{Lang} outperformed En-{Lang}. This suggests that providing instructions in English yields better results than in other languages. Interestingly, this contrasts with previous research (Afina et al., 2024b), which found that instructions in English led to better performance. While general QA tasks typically assess the model's knowledge, where multilingual knowledge can be effectively utilized, our explanation generation task, which requires the model to provide appropriate explanations using its knowledge, suggests that it struggles to deliver adequate explanations in languages other than English.

7 Conclusion

This study focused on artworks, which have a unique image and name regardless of the language, to evaluate the explanation generation abilities of LVLMs in multilingual contexts. We created datasets compiled from Wikipedia pages in ten languages without using machine translations to evaluate their abilities across multilingual languages. The results indicate that LVLMs perform optimally when input and output are both in English, while their performance declines when using languages other than English. Thus, our hypothesis, that "Vision Encoder needs to be learned in other languages as part of its pre-training," is correct, and might need to train Vision Transformer using multilingual data.

Limitations

Data Collection and Crawling Consistency

Our initial data collection was conducted through web crawling on June 30th, 2024. It is important to note that subsequent crawls may yield different results due to page updates, such as an increase in the number of pages or the addition of images. As a result, the data retrieved through repeated crawling may not consistently match the original dataset. This introduces a level of variability in the data, which must be considered when replicating or extending this research.

Necessity of Human Evaluation Across Multiple Languages

To validate the effectiveness and accuracy of LLMs, especially when dealing with complex and diverse linguistic features across multiple languages, human evaluation is indispensable. In this study, we conducted manual evaluations across ten languages. This step is crucial for assessing the model's real-world applicability and ensuring that automated evaluations do not overlook nuanced errors that only human evaluators can identify.

Coverage of Entity Knowledge in LVLMs

Artwork explanation requires an understanding of entities related to given artworks. However, similar to small-scale models (Kamigaito et al., 2023), LVLMs sometimes lack proper alignment of entity knowledge between LLMs and vision encoders. One possible way to address this issue is by leveraging external knowledge. For example, Vision RAG (Faysse et al., 2024) can search text from given images whereas it requires a retriever capable of interpreting artistic knowledge and LVLMs that can handle lengthy exemplars. Meanwhile, Multimodal Knowledge Graphs (KGs) (Chen et al., 2024) are concise yet sparse. Although PLM-based KG completion (KGC) (Yao et al., 2019) can potentially alleviate sparsity, its performance is uncertain due to the risk of data leakage (Sakai et al., 2024a). By contrast, traditional embedding-based KGC (Nickel et al., 2011) is robust and reliable, grounded in theoretical studies (Kamigaito and

Hayashi, 2021, 2022a,b; Feng et al., 2023, 2024), but its performance is generally lower than that of PLM-based KGC. Consequently, there is no decisive approach for integrating external knowledge into LVLMs for artwork explanation. The similar challenge arises in image review evaluation as reported by Hayashi et al. (2025).

Length of Generated Explanation

Our experimental results demonstrate that the length of the generated explanations can vary drastically across different LVLMs, potentially hindering fair comparisons. An instruction-based length control method (Juseon-Do et al., 2024) may mitigate this issue. However, such methods are primarily designed for summarization and their appropriateness for explanation generation is uncertain.

Ethical Considerations

Linguistic Considerations and Ethical Implications

In several languages, nouns are gendered, meaning they are classified as either masculine or feminine such as Spanish and Italian. For this study, we assumed that LLMs are capable of accurately distinguishing between these gendered forms. This assumption is crucial, as it reflects the model's ability to handle linguistic nuances, particularly in gendered languages. This raises ethical considerations, as any failure of the model to accurately represent gendered language could result in biased or incorrect outputs.

Wikipedia Resources among Ten Languages

Regarding Wikipedia pages, non-English versions are often less well-maintained, and whether entities are as well-organized as in English is debatable. In addition, Chinese Wikipedia contains a mix of traditional and simplified characters, which seems less standardized. In this study, since we crawled pages from Wikipedia and evaluated using their entities, it's possible that the correct answers are included in the outputs of LVLMs.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP23H03458.

References

Dubeyand Abhimanyu, Jauhriand Abhinav, Pandeyand Abhinav, Kadianand Abhishek, Al-Dahleand Ahmad, Letmanand Aiesha, Mathurand Akhil, Scheltenand Alan, Yangand Amy, Fanand Angela, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Putriand Rifki Afina, Haznitramaand Faiz Ghifari, Adhistaand Dea, and Ohand Alice. 2024a. Can llm generate culturally relevant commonsense qa data? case study in indonesian and sundanese. *arXiv preprint arXiv:2402.17302*.

Putriand Rifki Afina, Haznitramaand Faiz Ghifari, Adhistaand Dea, and Ohand Alice. 2024b. Can LLM generate culturally relevant commonsense QA data? case study in Indonesian and Sundanese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20571–20590, Miamiand Floridaand USA. Association for Computational Linguistics.

Dosovitskiyand Alexey, Beyerand Lucas, Kolesnikovand Alexander, Weissenbornand Dirk, Zhaiand Xiaohua, Unterthinerand Thomas, Dehghaniand Mostafa, Mindererand Matthias, Heigoldand Georg, Gellyand Sylvain, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

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. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Liand Bohao, Wangand Rui, Wangand Guangzhi, Geand Yuying, Geand Yixiao, and Shanand Ying. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

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. 2024. Knowledge graphs meet multi-modal learning: A comprehensive survey.

Linand Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Manuel Faysse, Hugues Sibille, Tony Wu, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models.

Xincan Feng, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2023. Model-based subsampling for knowledge graph completion. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 910–920, Nusa Dua, Bali. Association for Computational Linguistics.

Xincan Feng, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Unified interpretation of smoothing methods for negative sampling loss functions in knowledge graph embedding. In *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024)*, pages 78–98, Bangkok, Thailand. Association for Computational Linguistics.

Teamand Gemini, Aniland Rohan, Borgeaudand Sebastian, Wuand Yonghui, Alayracand Jean-Baptiste, Yuand Jiahui, Soricutand Radu, Schalkwykand Johan, Daiand Andrew M, Hauthand Anja, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Sasun Hambardzumyan, Abhinav Tuli, Levon Ghukasyan, Fariz Rahman, Hrant Topchyan, David Isayan, Mark McQuade, Mikayel Harutyunyan, Tatevik Hakobyan, Ivo Stranic, and Davit Buniatyan. 2022. Deep lake: a lakehouse for deep learning.

Liuand Haotian, Liand Chunyuan, Wuand Qingyang, and Leeand Yong Jae. 2023a. Visual instruction tuning.

Liuand Haotian, Liand Chunyuan, Liand Yuheng, Liand Bo, Zhangand Yuanhan, Shenand Sheng, and Leeand Yong Jae. 2024. Llava-next: Improved reasoningand ocrand and world knowledge.

Liuand Haotian, Liand Chunyuan, Liand Yuheng, and Leeand Yong Jae. 2023b. Improved baselines with visual instruction tuning.

Kazuki Hayashi, Kazuma Onishi, Toma Suzuki, Yusuke Ide, Seiji Gobara, Shigeki Saito, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2025. IRR: Image review ranking framework for evaluating vision-language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9939–9956, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Touvronand Hugo, Martinand Louis, Stoneand Kevin, Albertand Peter, Almahairiand Amjad, Babaeiand Yasmine, Bashlykovand Nikolay, Batraand Soumya, Bhargavaand Prajjwal, Bhosaleand Shruti, et al. 2023a. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Touvronand Hugo, Lavriland Thibaut, Izacardand Gautier, Martinetand Xavier, Lachauxand Marie-Anne, Lacroixand Timothée, Rozièreand Baptiste, Goyaland Naman, Hambroand Eric, Azharand Faisal, et al. 2023b. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Baiand Jinze, Baiand Shuai, Yangand Shusheng, Wangand Shijie, Tanand Sinan, Wangand Peng, Linand Junyang, Zhouand Chang, and Zhouand Jingren. 2023a. Qwen-vl: A versatile vision-language model for understandingand localizationand text readingand and beyond. *arXiv preprint arXiv:2308.12966*.

Baiand Jinze, Baiand Shuai, Yangand Shusheng, Wangand Shijie, Tanand Sinan, Wangand Peng, Linand Junyang, Zhouand Chang, and Zhouand Jingren. 2023b. Qwen-vl: A versatile vision-language model for understandingand localizationand text readingand and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Pfeifferand Jonas, Geigleand Gregor, Kamathand Aishwarya, Steitzand Jan-Martin O., Rothand Stefan, Vulićand Ivan, and Gurevychand Iryna. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublinand Ireland. Association for Computational Linguistics.

Liand Junnan, Liand Dongxu, Xiongand Caiming, and Hoiand Steven. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Liand Junnan, Liand Dongxu, Savareseand Silvio, and Hoiand Steven. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Juseon-Do, Jingun Kwon, Hidetaka Kamigaito, and Manabu Okumura. 2024. InstructCMP: Length control in sentence compression through instruction-based large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8980–8996, Bangkok, Thailand. Association for Computational Linguistics.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Heand Kaiming, Zhangand Xiangyu, Renand Shaoqing, and Sunand Jian. 2015. Deep residual learning for image recognition. arxiv e-prints. *arXiv preprint arXiv:1512.03385*, 10.

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.

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*, pages 498–517. Springer.

Papineniand Kishore, Roukosand Salim, Wardand Todd, and Zhuand Wei-Jing. 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.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. corr abs/1602.07332.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Wing Lian, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Slimorca: An open dataset of gpt-4 augmented flan reasoning tracesand with verification.

Chenand Lin, Liand Jinsong, Dongand Xiaoyi, Zhangand Pan, Heand Conghui, Wangand Jiaqi, Zhaoand Feng, and Linand Dahua. 2025. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. 2018. ivqa: Inverse visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8611–8619.

LLM-jp, :, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Akim Mousterou, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada, Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms.

Ouyangand Long, Wuand Jeffrey, Jiangand Xu, Almeidaand Diogo, Wainwrightand Carroll, Mishkinand Pamela, Zhangand Chong, Agarwaland Sandhini, Slamaand Katarina, Rayand Alex, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems*, 35:2507–2521.

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. *arXiv preprint arXiv:2110.13214*.

Nguyenand Ngan Luu-Thuy, Nguyenand Nghia Hieu, Voand Duong TD, Tranand Khanh Quoc, and Van Nguyenand Kiet. 2023. Vlsp2022-evjvqa challenge: Multilingual visual question answering. *arXiv preprint arXiv:2302.11752*.

Reidand Machel, Savinovand Nikolay, Teplyashinand Denis, Lepikhinand Dmitry, Lillicrapand Timothy, Alayracand Jean-baptiste, Soricutand Radu, Lazaridouand Angeliki, Firatand Orhan, Schrittwieserand Julian, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Abdinand Marah, Jacobsand Sam Ade, Awanand Ammar Ahmad, Anejaand Jyoti, Awadallahand Ahmed, Awadallaand Hany, Bachand Nguyen, Bahreeand Amit, Bakhtiariand Arash, Behland Harkirat, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

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. IEEE.

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.

Sidorovand Oleksii, Huand Ronghang, Rohrbachand Marcus, and Singhand Amanpreet. 2020. Textcaps: a dataset for image captioningwith reading comprehension. *European Conference on Computer Vision*.

Shintaro Ozaki, Kazuki Hayashi, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024a. Bqa: Body language question answering dataset for video large language models.

Shintaro Ozaki, Yuta Kato, Siyuan Feng, Masayo Tomita, Kazuki Hayashi, Ryoma Obara, Masafumi Oyamada, Katsuhiko Hayashi, Hidetaka Kamigaito, and Taro Watanabe. 2024b. Understanding the impact of confidence in retrieval augmented generation: A case study in the medical domain.

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 preprint arXiv:2306.14824*.

Yeand Qinghao, Xuand Haiyang, Yeand Jiabo, Yanand Ming, Huand Anwen, Liuand Haowei, Qianand Qi, Zhangand Ji, and Huangand Fei. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024a. Does pre-trained language model actually infer unseen links in knowledge graph completion? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8091–8106, Mexico City, Mexico. Association for Computational Linguistics.

Yusuke Sakai, Adam Nohejl, Jiangnan Hang, Hidetaka Kamigaito, and Taro Watanabe. 2024b. Toward the evaluation of large language models considering score variance across instruction templates. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 499–529, Miami, Florida, US. Association for Computational Linguistics.

Takehiro Sato, Shintaro Ozaki, and Daisaku Yokoyama. 2024. An implementation of werewolf agent that does not truly trust LLMs. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 58–67, Tokyo, Japan. Association for Computational Linguistics.

Samir Yitzhak Gadreand Gabriel Ilharcoand Alex Fangand Jonathan Hayaseand Georgios Smyrnisand Thao Nguyenand Ryan Martenand Mitchell Wortsmanand Dhruba Ghoshand Jieyu Zhangand Eyal Orgadand Rahim Entezariand Giannis Darasand Sarah Prattand Vivek Ramanujanand Yonatan Bittonand Kalyani Maratheand Stephen Mussmannand Richard Vencuand Mehdi Chertiand Ranjay Krishnaand Pang Wei Kohand Olga Saukhand Alexander Ratnerand Shuran Songand Hannaneh Hajishirziand Ali Farhadiand Romain Beaumontand Sewoong Ohand Alex Dimakisand Jenia Jitsevand Yair Carmonand Vaishaal Shankarand Ludwig Schmidt. 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. In *European conference on computer vision*, pages 146–162. Springer.

Changpinyoand Soravit, Xueand Linting, Yaromand Michal, Thapliyaland Ashish, Szpektorand Idan, Amelotand Julien, Chenand Xi, and Soricutand Radu. 2023. MaXM: Towards multilingual visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682, Singapore. Association for Computational Linguistics.

Changpinyoand Soravit, Sharmaand Piyush, Dingand Nan, and Soricutand Radu. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Wolfand Thomas, Debutand Lysandre, Sanhand Victor, Chaumondand Julien, Delangueand Clement, Moiand Anthony, Cistacand Pierric, Raultand Tim, Loufand Remi, Funtowiczand Morgan, Davisonand Joe, Shleiferand Sam, von Platenand Patrick, Maand Clara, Jerniteand Yacine, Pluand Julien, Xuand Canwen, Le Scaoand Teven, Guggerand Sylvain, Drameand Mariama, Lhoestand Quentin, and Rushand Alexander. 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.

Zhangand Tianyi, Kishoreand Varsha, Wuand Felix, Weinbergerand Kilian Q, and Artziand Yoav. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Dettmersand Tim, Lewisand Mike, Belkadaand Younes, and Zettlemoyerand Luke. 2022. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332.

Chenand Ting, Kornblithand Simon, Norouziand Mohammad, and Hintonand Geoffrey. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Brownand Tom, Mannand Benjamin, Ryderand Nick, Subbiahand Melanie, Kaplanand Jared D, Dhariwaland Prafulla, Neelakantanand Arvind, Shyamand Pranav, Sastryand Girish, Askelland Amanda, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ordonezand Vicente, Kulkarniand Girish, and Bergand Tamara. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associatesand Inc.

Dongand Xiaoyi, Zhangand Pan, Zangand Yuhang, Caoand Yuhang, Wangand Bin, Ouyangand Linke, Weiand Xilin, Zhangand Songyang, Duanand Haodong, Caoand Maosong, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Chenand Xinlei, Fangand Hao, Linand Tsung-Yi, Vedantamand Ramakrishna, Guptaand Saurabh, Dollárand Piotr, and Zitnickand C Lawrence. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion.

Liuand Yuan, Duanand Haodong, Zhangand Yuanhan, Liand Bo, Zhangand Songyang, Zhaoand Wangbo, Yuanand Yike, Wangand Jiaqi, Heand Conghui, Liuand Ziwei, et al. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Sakaiand Yusuke, Kamigaitoand Hidetaka, and Watanabeand Taro. 2024. mcsqa: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. *arXiv preprint arXiv:2406.04215*.

Sina Zarrieß and David Schlangen. 2018. Decoding strategies for neural referring expression generation. *t*, pages 503–512.

A Appendix

A.1 Inference and LoRA Tuning Settings

Inference setting

In this study, as far as inference which needs to use GPUs, all experiments were conducted on a single NVIDIA RTX A6000 GPU and NVIDIA A100-SXM4-40GB, with 8-bit quantization utilized for model generation. However, there is no InternLM-XComposer-2 with 8-bit, this 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.

LoRA Tuning setting

We conducted LoRA (Hu et al., 2022) Tuning with two models: LLaVA-NeXT and Qwen-VL. Both were trained using two NVIDIA A100-SXM4-40GB GPUs. Detailed parameters are provided in Table 5 and Table 6.

B Explanation Generation Abilities from Other Languages to English ({Lang}-{En})

When considering output in English from other languages, we found this method also performs less abilities. This suggests that LVLMs have relatively less training data in languages other than English, and they may not properly understand instructions given in other languages. Thus, it is difficult to say that the integrated learning of LLMs and Vision Encoder work properly.

C Details of experimental settings

Model	Base Model	HuggingFace Name
mPLUG-Owl2	LLaMA2-7B	MAGAer13/mplug-owl2-llama2-7b
Qwen-VL-Chat	Qwen	Qwen/Qwen-VL-Chat
LLaVA-NeXT	LLaMA3-8B	lmms-lab/llama3-llava-next-8b
Phi-3	Phi-3-Vision-128K-Instruct	microsoft/Phi-3-vision-128k-instruct
XComposer2	internlm-xcomposer2-7B	internlm/internlm-xcomposer2-7B

Hyper Parameter	Value
torch_dtype	bfloat16
seed	42
max length	2,048
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 5: The hyper-parameters of Qwen-VL used in the experiment, and others, were set to default settings. The implementation used Transformers (Thomas et al., 2020) and bitsandbytes (Tim et al., 2022).

Hyper Parameter	Value
seed	42
max length	2048
lora enable	True
learning rate	2e-5
warmup ratio	0.05
lora r	16
lora alpha	32
torch_dtype	float16

Table 6: The hyper-parameters of LLaVA-NeXT used in the experiment, and others were also set to default settings.

D Details of Creating Datasets or Training Data

D.1 How to Choose Ten Languages?

We selected ten languages with the highest number of articles from the statistics of all language versions of Wikipedia[5]. Of the top 10 prefectures, Cebuano, Egyptian dialects of Arabic, and Polish were deemed difficult to identify by sampling during the evaluation, so we added the runners-up, Chinese and Japanese.

D.2 How to Split Train, Valid, and Test Data in English?

For English, a language rich resource, we split the data into train, valid, and test data using six metrics proposed by Hayashi et al. (2024) (six metrics: page views, number of links, number of edits, number of references, number of language versions, and article length.) were used in this study as well, and the data were divided equally considering famous artworks. All data included in

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

Language	A-10	A-5	Full			#NUM in A-10			#NUM in A-5			#NUM in Full
	Test	Test	Train	Valid	Test	En-{L}	{L}-En	{L}-{L}	En-{L}	{L}-En	{L}-{L}	En-{L}
English	28	306	6,413	2,138	2,138	864	–	–	5,924	–	–	29,064
French	28	306	–	–	2,707	1,780	1,180	856	11,020	6,184	6,176	57,916
Spanish	28	306	–	–	2,096	964	1,176	668	5,744	6,056	4,504	40,960
Italian	28	306	–	–	1,977	1,092	1,244	596	6,396	7,032	4,900	39,092
Japanese	28	306	–	–	1,312	1,012	1,244	572	6,156	6,928	4,720	23,760
Russian	28	–	–	–	1,589	668	1,184	604	–	–	–	27,572
German	28	–	–	–	962	732	1,208	1,132	–	–	–	23,792
Dutch	28	–	–	–	789	556	1,284	468	–	–	–	13,940
Swedish	28	–	–	–	774	912	1,164	444	–	–	–	7,228
Chinese	28	–	–	–	738	1,092	1,304	612	–	–	–	8,624
ALL	280	1,530	-	-	15,082	9,672	9,704	5,952	35,240	26,200	20,300	271,948

Table 7: The number of each language data in Alignment-10, Alignment-5, and Full task, split by train, valid, and test sets. We split train, valid and test sets only English due to the number of data in English. #NUM represents the total number of inferences, and "A" indicates Alignment, showing the number of aligned artworks across languages. Additionally, {L} is an abbreviation for {Lang}, representing nine languages other than English.

the alignment were used as test data so that data used in the alignment task were not included in the train. We described the number of all data in Table 7.

D.3 License

In our study, we created a dataset from Wikipedia articles regarding artworks. 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. Thus, our data does not infringe upon the licenses.

E Evaluation Metrics Formulation

This section describes on the evaluation metrics used in Section 5 using mathematical expressions (Hayashi et al., 2024; Ozaki et al., 2024b). 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 $\text{Entity}(\cdot)$ 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:

$$EC(G, R) = Cov(G, R) \quad (1)$$

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:

$$EF_1 = \frac{2 \times P \times R}{P + R} \quad (2)$$

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

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

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:

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

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

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

and the 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.[6]

F Filtered Sections

The following section was filtered in this study. Approximately 30 instances from the Alignment-10 task were reviewed, and sections without informative content.

English

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

Japanese

外部リンク, 参考文献, 関連項目, 脚注, 出典, ギャラリー, バージョン, 注釈, 関連する作品

Italian

Collegamenti esterni, Altri progetti, Bibliografia, Note, Omaggi, Voci correlate, Bibliografia, Musica, Fumetti, Letteratura, Filmografia, Nella cultura di massa, Altri progetti, Galleria d'immagini, Curiosità, Calendario

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

French

Liens externes, Articles connexes, Bibliographie et ressources en ligne, Annexes, Notes et références, Divers, Littérature, Peinture et sculpture, Déclinaisons et détournements, Bases de données et dictionnaires, Italien, Français, Ouvrages, Articles, Bibliographie, Théâtre, Cinéma, Article connexe, Annexe, Notes et référence, Voir aussi, Divers, Pour approfondir, Versions, Références, Sources secondaires, Sources originales, Références de l'expression dans l'art, Ouvrages, Ailleurs, Notes, Films, Dans la culture, Postérité, Données techniques, Galerie, Historique

Spanish

Enlaces externos, Bibliografía, Referencias, Fuentes, Enlaces externos, Bibliografía, Véase también, Notas, Información, Galería, Galería de imágenes, Filmografía

Chinese (Traditional)

外部連結, 延伸, 参考文献, 參考文獻, 參見, 参見, 書目, 注与参考文献, 來源, 擴展讀, 參考来源, 外部接, 延伸, 引用, 注, 參考資料, 參考料, 相關條目, 參考來源, 參見條目, 其他事項, 參考, 註解, 媒體, 紀錄片, 書籍, 近似作品, 相關作品, 德文, 注, 擴展讀, 吉米·威士的声明

Chinese (Simplified)

外部链接, 延伸阅读, 参考文献, 参见, 注释与参考文献, 来源, 扩展阅读, 参考来源, 引用, 注释, 参考资料, 相关条目, 参见条目, 其他事项, 参考, 近似作品, 媒体, 纪录片, 书籍, 注释, 吉米·威尔士的声明

Swedish

Noter, Referenser, Se även, Externa länkar, Allmänna källor, Galleri, Källor, Bilder, Kalenderfunktionen, Relaterade målningar

Dutch

Zie ook, Literatuur, Externe links, Bewerkingen, Andere, Latere edities, Trivia, Zie ook, Galerij, Originele gietingen, Stanza dell'incendio del Borgo, Stanza della Segnatura, Noten, Literatuur en bronnen

Russian

Ссылки, Примечания, См. также, Документалистика, Литература, Источники, Отражение в искусстве

German

Anmerkungen, Weblinks, Literatur, Anmerkungen und Einzelnachweise, Einzelbelege, Einzelnachweise, Chronologie, Quellen, Übersicht, Literatur (Auswahl), Siehe auch, Rezeption, Dokumentarfilme, Ausstellungen, Siehe auch

G Instruction to Native Speakers

We asked native speaker to prepare the instruction to check if the above template is equal in difficulty compared to the English text.

Instruction

What we research
We are conducting a study to measure LLMs' ability to understand the arts. Previous studies have been done only for English, and we are now trying to extend and validate it for multiple languages.

The text presented has been translated from English into your language using DeepL.

I want you to make sure that the sentence you translate has the same meaning as the English sentence.
The time I assume will not take more than 5 minutes and that's about OK for a check. I also use back translation to check it, so I believe it is not that broken.

Keep in mind

- My final goal is to have the sentences corrected to be as natural as English sentences.
- Depending on {title} and {section}, and in some countries, you may need to be concerned about masculine and feminine nouns. If that is the case, choose whichever you type into the LLM in your native language (i.e., the more natural one).
- Please do not change the entire text.
- Changing, deleting or adding words is acceptable.

Examples of {title} and {section}, {subsection} and {subsubsection}
We use Wikipedia for our research.
Here is one of the example: `https://en.wikipedia.org/wiki/Mona_Lisa`
In this case, {title} will contain "Mona Lisa".
In addition, {section} contains "Description", "History", and so on.
{subsection} refers to a smaller frame within {section}, such as "Creation and date".

↓ Below is the text I would like you to review.
({lang}_temp1_sec is translated from en_temp1_sec using DeepL) ({lang}_temp2_subsec is translated from en_temp2_subsec using DeepL)

#English (source)
This sentence is a sample.

Your native language (target I translated from DeepL.)
This sentence is a sample.

H Other Results and Visualizations

Input	Output	LVLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
			exact	partial		n=0	n=1	n=2	n=∞	
En	En	LLaVA-NeXT	17.66	26.05	18.55	1.31	1.23	1.20	1.20	242
		LLaVA-NeXT (FT)	17.92	23.65	19.20	**5.67**	5.66	5.63	5.60	81
		Qwen-VL	14.60	21.51	15.39	1.36	1.33	1.28	1.27	110
		Qwen-VL (FT)	**20.09**	**26.27**	**20.84**	5.67	**5.78**	**5.77**	**5.72**	171
		mPLUG-Owl2	14.41	21.96	15.71	1.27	1.17	1.14	1.10	121
En	Es	LLaVA-NeXT	**10.40 (-7.26)**	**16.05 (-10.00)**	**10.86 (-7.69)**	**0.79 (-0.52)**	**0.78 (-0.45)**	**0.83 (-0.37)**	**0.83 (-0.37)**	181
		LLaVA-NeXT (FT)	4.96 (-12.96)	8.42 (-15.23)	5.40 (-13.80)	0.56 (-5.10)	0.57 (-5.10)	0.58 (-5.04)	0.57 (-5.02)	90
		Qwen-VL	8.11 (-6.49)	13.18 (-8.33)	8.66 (-6.73)	0.53 (-0.83)	0.50 (-0.83)	0.52 (-0.76)	0.51 (-0.76)	103
		Qwen-VL (FT)	4.23 (-15.86)	8.47 (-17.80)	4.66 (-16.17)	0.23 (-5.43)	0.23 (-5.56)	0.24 (-5.53)	0.24 (-5.48)	195
		mPLUG-Owl2	7.26 (-7.14)	12.13 (-9.83)	7.55 (-8.16)	0.45 (-0.82)	0.49 (-0.68)	0.52 (-0.63)	0.51 (-0.59)	100
En	Fr	LLaVA-NeXT	**9.71 (-7.95)**	**16.17 (-9.88)**	**9.49 (-9.06)**	**0.57 (-0.74)**	**0.57 (-0.66)**	**0.57 (-0.63)**	**0.55 (-0.64)**	168
		LLaVA-NeXT (FT)	7.02 (-10.90)	10.37 (-13.29)	7.60 (-11.60)	**0.84 (-4.83)**	**0.84 (-4.82)**	**0.82 (-4.81)**	**0.81 (-4.79)**	60
		Qwen-VL	7.64 (-6.96)	12.82 (-8.68)	7.71 (-7.68)	0.51 (-0.85)	0.46 (-0.87)	0.45 (-0.83)	0.43 (-0.83)	86
		Qwen-VL (FT)	6.42 (-13.68)	11.17 (-15.10)	6.88 (-13.95)	0.43 (-5.24)	0.56 (-5.22)	0.55 (-5.22)	0.54 (-5.18)	155
		mPLUG-Owl2	6.99 (-7.42)	12.55 (-9.41)	6.91 (-8.79)	0.41 (-0.86)	0.38 (-0.79)	0.37 (-0.77)	0.35 (-0.75)	95
En	De	LLaVA-NeXT	**10.32 (-7.35)**	**13.84 (-12.21)**	**12.28 (-6.27)**	**0.90 (-0.41)**	**0.88 (-0.34)**	**0.88 (-0.32)**	**0.86 (-0.34)**	161
		LLaVA-NeXT (FT)	5.52 (-12.40)	7.80 (-15.86)	5.93 (-13.26)	0.52 (-5.15)	0.48 (-5.19)	0.46 (-5.17)	0.45 (-5.15)	75
		Qwen-VL	7.75 (-6.85)	10.60 (-10.91)	8.69 (-6.69)	0.63 (-0.73)	0.59 (-0.74)	0.58 (-0.70)	0.56 (-0.71)	99
		Qwen-VL (FT)	4.79 (-15.30)	7.40 (-18.87)	5.17 (-15.67)	0.23 (-5.44)	0.25 (-5.53)	0.24 (-5.53)	0.24 (-5.48)	177
		mPLUG-Owl2	6.87 (-7.53)	9.66 (-12.30)	7.69 (-8.01)	0.60 (-0.67)	0.54 (-0.63)	0.53 (-0.61)	0.50 (-0.60)	91
En	It	LLaVA-NeXT	**9.57 (-8.10)**	**16.52 (-9.53)**	**10.72 (-7.83)**	0.72 (-0.60)	0.72 (-0.50)	0.74 (-0.46)	0.72 (-0.47)	168
		LLaVA-NeXT (FT)	6.21 (-11.71)	9.51 (-14.15)	7.59 (-11.61)	**0.79 (-4.88)**	**0.86 (-4.80)**	**0.85 (-4.77)**	**0.85 (-4.74)**	87
		Qwen-VL	7.08 (-7.52)	12.73 (-8.77)	8.26 (-7.13)	0.34 (-1.02)	0.38 (-0.95)	0.38 (-0.90)	0.38 (-0.88)	112
		Qwen-VL (FT)	6.08 (-14.01)	10.10 (-16.17)	7.39 (-13.44)	0.49 (-5.17)	0.58 (-5.20)	0.59 (-5.19)	0.59 (-5.13)	187
		mPLUG-Owl2	6.54 (-7.86)	12.20 (-9.76)	7.44 (-8.27)	0.42 (-0.85)	0.40 (-0.77)	0.39 (-0.75)	0.39 (-0.71)	102
En	Nl	LLaVA-NeXT	**7.91 (-9.76)**	**13.25 (-12.80)**	8.63 (-9.92)	0.31 (-1.01)	0.44 (-0.79)	0.42 (-0.78)	0.43 (-0.77)	175
		LLaVA-NeXT (FT)	7.89 (-10.03)	11.66 (-12.00)	**8.81 (-10.39)**	**1.22 (-4.44)**	**1.13 (-4.53)**	**1.12 (-4.51)**	**1.11 (-4.49)**	102
		Qwen-VL	7.41 (-7.19)	12.33 (-9.18)	7.93 (-7.46)	0.35 (-1.01)	0.49 (-0.84)	0.50 (-0.78)	0.53 (-0.74)	137
		Qwen-VL (FT)	6.67 (-13.42)	10.07 (-16.21)	7.67 (-13.16)	0.68 (-4.98)	0.73 (-5.05)	0.70 (-5.08)	0.71 (-5.01)	166
		mPLUG-Owl2	4.61 (-9.80)	8.96 (-13.00)	4.84 (-10.87)	0.20 (-1.06)	0.26 (-0.91)	0.25 (-0.89)	0.25 (-0.85)	106
En	Sv	LLaVA-NeXT	**13.08 (-4.59)**	**17.19 (-8.85)**	**12.38 (-6.18)**	**0.89 (-0.42)**	**0.82 (-0.41)**	**0.82 (-0.39)**	**0.75 (-0.44)**	172
		LLaVA-NeXT (FT)	9.44 (-8.47)	12.79 (-10.87)	9.62 (-9.58)	0.73 (-4.94)	0.64 (-5.03)	0.60 (-5.03)	0.58 (-5.01)	94
		Qwen-VL	10.59 (-4.01)	14.72 (-6.79)	10.75 (-4.64)	0.58 (-0.78)	0.61 (-0.72)	0.66 (-0.62)	0.61 (-0.66)	124
		Qwen-VL (FT)	9.47 (-10.62)	13.20 (-13.07)	9.93 (-10.91)	0.72 (-4.95)	0.65 (-5.13)	0.63 (-5.14)	0.58 (-5.14)	155
		mPLUG-Owl2	9.37 (-5.03)	12.82 (-9.14)	8.53 (-7.17)	0.40 (-0.86)	0.36 (-0.81)	0.36 (-0.78)	0.33 (-0.77)	79
En	Ru	LLaVA-NeXT	**7.86 (-9.81)**	**10.75 (-15.29)**	**6.39 (-12.16)**	**0.22 (-1.09)**	**0.26 (-0.97)**	**0.28 (-0.92)**	**0.28 (-0.92)**	203
		LLaVA-NeXT (FT)	0.42 (-17.50)	1.51 (-22.14)	0.31 (-18.89)	0.01 (-5.66)	0.01 (-5.65)	0.01 (-5.62)	0.01 (-5.58)	72
		Qwen-VL	3.05 (-11.55)	4.81 (-16.69)	2.35 (-13.04)	0.05 (-1.31)	0.07 (-1.26)	0.08 (-1.20)	0.08 (-1.18)	112
		Qwen-VL (FT)	0.15 (-19.94)	1.09 (-25.19)	0.09 (-20.74)	0.00 (-5.67)	0.00 (-5.78)	0.00 (-5.77)	0.00 (-5.72)	203
		mPLUG-Owl2	3.69 (-10.71)	5.33 (-16.64)	2.83 (-12.88)	0.11 (-1.16)	0.10 (-1.07)	0.09 (-1.05)	0.10 (-1.00)	107
En	Ja	LLaVA-NeXT	**8.65 (-9.01)**	**8.70 (-17.35)**	**12.34 (-6.21)**	**0.44 (-0.87)**	**0.44 (-0.79)**	**0.44 (-0.76)**	**0.44 (-0.76)**	213
		LLaVA-NeXT (FT)	0.46 (-17.45)	0.61 (-23.04)	0.51 (-18.69)	0.02 (-5.65)	0.02 (-5.65)	0.02 (-5.61)	0.02 (-5.58)	67
		Qwen-VL	3.10 (-11.50)	3.16 (-18.35)	4.37 (-11.02)	0.12 (-1.24)	0.12 (-1.21)	0.12 (-1.16)	0.12 (-1.14)	127
		Qwen-VL (FT)	0.21 (-19.88)	0.46 (-25.82)	0.12 (-20.72)	0.00 (-5.67)	0.00 (-5.78)	0.00 (-5.77)	0.00 (-5.72)	152
		mPLUG-Owl2	4.00 (-10.40)	4.06 (-17.90)	5.39 (-10.32)	0.25 (-1.01)	0.25 (-0.92)	0.25 (-0.89)	0.25 (-0.85)	104
En	Zh	LLaVA-NeXT	**10.81 (-6.86)**	**10.90 (-15.15)**	**13.00 (-5.56)**	0.60 (-0.71)	0.60 (-0.62)	0.60 (-0.60)	0.60 (-0.59)	220
		LLaVA-NeXT (FT)	0.64 (-17.27)	0.89 (-22.76)	0.75 (-18.45)	0.08 (-5.59)	0.08 (-5.59)	0.08 (-5.55)	0.08 (-5.52)	71
		Qwen-VL	8.60 (-6.00)	8.65 (-12.85)	10.34 (-5.05)	**0.80 (-0.56)**	**0.79 (-0.54)**	**0.79 (-0.49)**	**0.79 (-0.47)**	133
		Qwen-VL (FT)	0.35 (-19.74)	0.64 (-25.63)	0.27 (-20.57)	0.01 (-5.66)	0.01 (-5.77)	0.01 (-5.77)	0.01 (-5.71)	155
		mPLUG-Owl2	4.99 (-9.42)	5.04 (-16.92)	6.08 (-9.62)	0.52 (-0.75)	0.52 (-0.65)	0.52 (-0.63)	0.52 (-0.58)	107

Table 8: Results of LVLMs in Full Task. Bold fonts indicate the best score for that language combination. This result shows that no matter how much the amount of data is increased, the best performance is achieved by having instructions given and output in English. The values are noted next to the output of the difference by the same model in the method with instruction and output in English ({En}-{En}). Red indicates a higher value than that method; blue indicates a lower value.

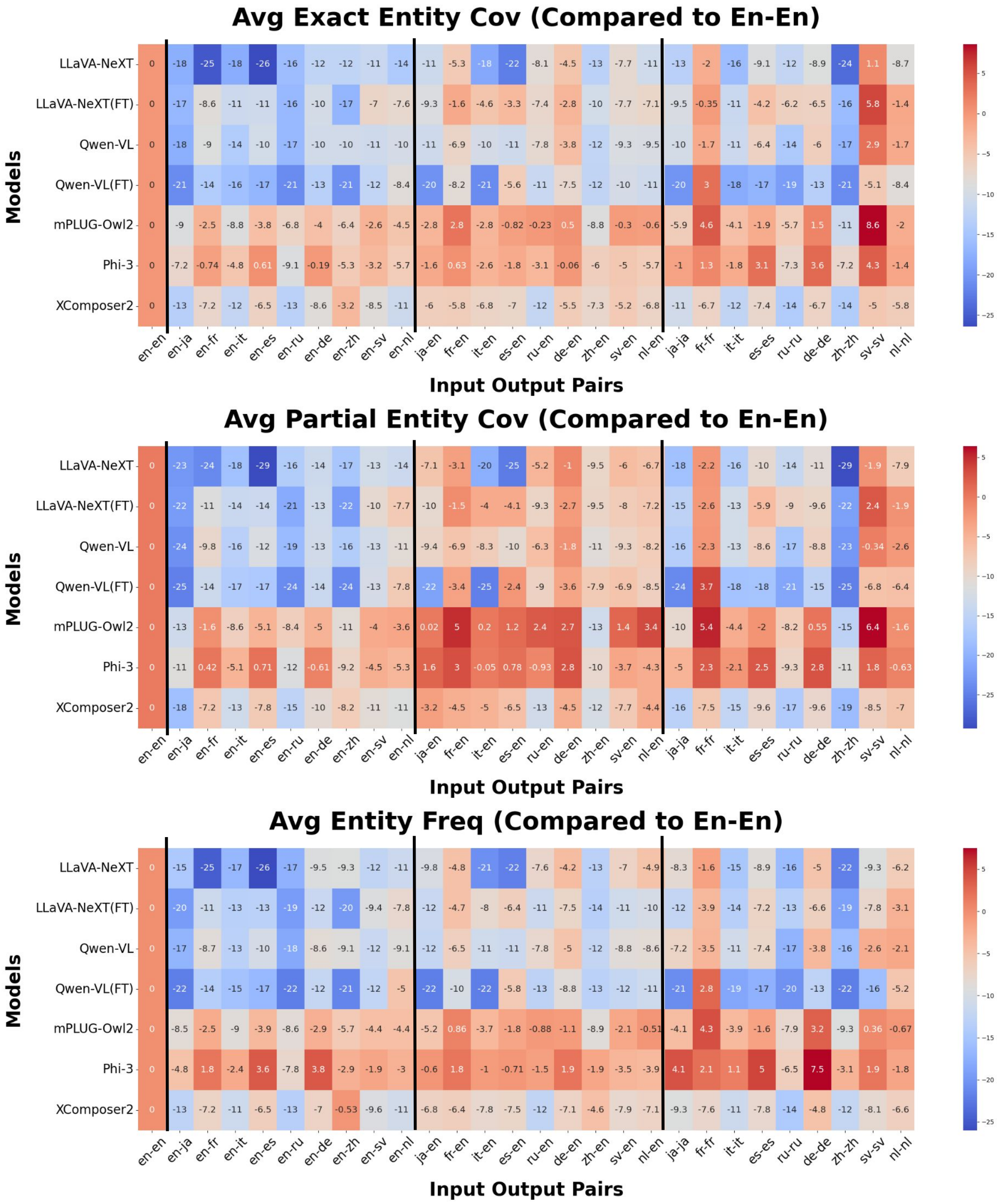

Figure 4: Visualization of Alignment-10 results in a heat map. We made the visualization based on when we had LVLMs give instructions and output in English.

Input	Output	LVLM	Entity Cov. exact	Entity Cov. partial	Entity F1	Entity Cooccurrence n=0	Entity Cooccurrence n=1	Entity Cooccurrence n=2	Entity Cooccurrence n=∞	Avg. Length
En	En	LLaVA-NeXT	**22.58**	**30.30**	**25.19**	3.15	2.30	2.40	2.44	259
		LLaVA-NeXT (FT)	13.80	18.73	14.57	3.68	3.55	3.50	3.42	86
		Qwen-VL	19.74	26.57	20.79	2.14	2.18	2.20	1.99	178
		Qwen-VL (FT)	19.23	25.31	19.16	**5.70**	**4.93**	**4.88**	**4.78**	237
		mPLUG-Owl2	18.76	26.05	20.90	2.74	1.99	1.93	1.83	158
		Phi-3	12.12	18.55	12.05	1.68	0.97	1.01	1.00	133
		XComposer2	21.40	28.90	22.96	3.25	2.62	2.54	2.39	242
En	Es	LLaVA-NeXT	5.76 (-16.82)	8.21 (-22.09)	7.39 (-17.80)	0.38 (-2.77)	0.34 (-1.96)	0.35 (-2.05)	0.32 (-2.12)	160
		LLaVA-NeXT (FT)	4.76 (-9.04)	8.63 (-10.10)	5.42 (-9.15)	0.50 (-3.18)	0.54 (-3.01)	0.52 (-2.97)	0.51 (-2.91)	83
		Qwen-VL	**8.52 (-11.23)**	**14.11 (-12.45)**	**8.82 (-11.97)**	**0.57 (-1.57)**	**0.61 (-1.57)**	**0.62 (-1.58)**	**0.61 (-1.38)**	118
		Qwen-VL (FT)	4.33 (-14.90)	8.63 (-16.68)	5.23 (-13.92)	0.18 (-5.52)	0.27 (-4.65)	0.28 (-4.60)	0.27 (-4.51)	185
		mPLUG-Owl2	7.30 (-11.46)	12.63 (-13.42)	7.19 (-13.71)	0.52 (-2.23)	0.50 (-1.48)	0.48 (-1.46)	0.46 (-1.37)	104
		Phi-3	7.28 (-4.84)	12.71 (-5.84)	7.14 (-4.90)	0.52 (-1.16)	0.56 (-0.41)	0.59 (-0.41)	0.55 (-0.45)	142
		XComposer2	7.16 (-14.23)	12.28 (-16.62)	7.07 (-15.88)	0.43 (-2.82)	0.43 (-2.20)	0.41 (-2.13)	0.40 (-1.99)	105
Es	En	LLaVA-NeXT	**29.15 (+6.56)**	**33.34 (+3.04)**	**29.88 (+4.69)**	0.35 (-2.79)	0.32 (-1.98)	**2.05 (-0.36)**	**2.01 (-0.43)**	220
		LLaVA-NeXT (FT)	10.32 (-3.48)	14.70 (-4.03)	11.08 (-3.49)	1.24 (-2.45)	1.29 (-2.26)	1.27 (-2.23)	1.25 (-2.17)	116
		Qwen-VL	10.95 (-8.79)	15.95 (-10.62)	11.99 (-8.80)	0.78 (-1.36)	0.82 (-1.36)	0.81 (-1.38)	0.74 (-1.25)	58
		Qwen-VL (FT)	15.18 (-4.05)	21.00 (-4.31)	15.98 (-3.18)	**1.84 (-3.86)**	**2.01 (-2.91)**	2.01 (-2.87)	1.93 (-2.85)	204
		mPLUG-Owl2	10.22 (-8.55)	15.39 (-10.66)	10.80 (-10.10)	0.97 (-1.78)	0.80 (-1.18)	0.78 (-1.15)	0.74 (-1.09)	56
		Phi-3	6.85 (-5.27)	12.14 (-6.41)	6.71 (-5.34)	0.52 (-1.16)	0.46 (-0.51)	0.49 (-0.52)	0.45 (-0.55)	95
		XComposer2	8.65 (-12.75)	13.28 (-15.62)	9.11 (-13.85)	0.51 (-2.75)	0.53 (-2.09)	0.50 (-2.04)	0.44 (-1.95)	71
Es	Es	LLaVA-NeXT	**12.77 (-9.81)**	**18.26 (-12.04)**	**10.97 (-14.22)**	**1.29 (-1.85)**	**1.22 (-1.08)**	**1.22 (-1.18)**	**1.08 (-1.35)**	203
		LLaVA-NeXT (FT)	9.47 (-4.34)	14.40 (-4.33)	8.14 (-6.43)	0.83 (-2.86)	0.79 (-2.76)	0.85 (-2.64)	0.70 (-2.72)	141
		Qwen-VL	10.40 (-9.34)	15.42 (-11.14)	9.70 (-11.09)	0.99 (-1.15)	0.93 (-1.25)	0.94 (-1.26)	1.07 (-0.92)	149
		Qwen-VL (FT)	3.12 (-16.11)	7.08 (-18.23)	3.40 (-15.76)	0.13 (-5.58)	0.13 (-4.80)	0.13 (-4.75)	0.12 (-4.66)	299
		mPLUG-Owl2	8.55 (-10.22)	13.55 (-12.49)	7.69 (-13.22)	0.89 (-1.86)	0.73 (-1.26)	0.76 (-1.17)	0.75 (-1.08)	132
		Phi-3	9.24 (-2.88)	14.38 (-4.18)	7.84 (-4.20)	0.63 (-1.05)	0.53 (-0.44)	0.55 (-0.46)	0.52 (-0.48)	198
		XComposer2	5.83 (-15.57)	9.87 (-19.03)	5.36 (-17.60)	0.51 (-2.75)	0.45 (-2.17)	0.47 (-2.07)	0.45 (-1.94)	88
En	Fr	LLaVA-NeXT	6.85 (-15.73)	12.94 (-17.37)	7.03 (-18.16)	0.22 (-2.93)	0.16 (-2.14)	0.17 (-2.24)	0.17 (-2.27)	181
		LLaVA-NeXT (FT)	6.06 (-7.74)	9.54 (-9.20)	6.36 (-8.22)	**0.72 (-2.96)**	**0.83 (-2.72)**	**0.83 (-2.67)**	**0.83 (-2.59)**	72
		Qwen-VL	**8.18 (-11.57)**	**13.52 (-13.05)**	**8.18 (-12.60)**	0.63 (-1.51)	0.61 (-1.57)	0.61 (-1.59)	0.60 (-1.39)	108
		Qwen-VL (FT)	5.91 (-13.32)	10.40 (-14.91)	6.40 (-12.76)	0.28 (-5.42)	0.43 (-4.49)	0.43 (-4.45)	0.42 (-4.36)	177
		mPLUG-Owl2	6.94 (-11.82)	12.38 (-13.67)	6.70 (-14.20)	0.55 (-2.19)	0.44 (-1.54)	0.42 (-1.51)	0.41 (-1.42)	109
		Phi-3	5.76 (-6.36)	10.92 (-7.63)	5.66 (-6.39)	0.43 (-1.25)	0.34 (-0.63)	0.34 (-0.66)	0.33 (-0.67)	149
		XComposer2	5.84 (-15.56)	10.78 (-18.12)	5.58 (-17.38)	0.42 (-2.84)	0.33 (-2.30)	0.31 (-2.23)	0.30 (-2.09)	79
Fr	En	LLaVA-NeXT	3.57 (-19.01)	5.33 (-24.97)	3.27 (-21.93)	0.00 (-3.15)	0.00 (-2.30)	0.00 (-2.40)	0.00 (-2.44)	162
		LLaVA-NeXT (FT)	10.60 (-3.20)	15.37 (-3.36)	11.16 (-3.41)	1.05 (-2.63)	1.00 (-2.55)	0.96 (-2.54)	0.89 (-2.53)	90
		Qwen-VL	**13.56 (-6.18)**	**18.70 (-7.87)**	**14.93 (-5.86)**	1.09 (-1.05)	1.15 (-1.03)	**1.18 (-1.02)**	**1.13 (-0.86)**	141
		Qwen-VL (FT)	11.20 (-8.03)	16.94 (-8.37)	11.23 (-7.93)	0.71 (-4.99)	0.78 (-4.14)	0.78 (-4.11)	0.76 (-4.02)	266
		mPLUG-Owl2	12.99 (-5.77)	18.51 (-7.54)	13.64 (-7.27)	**1.11 (-1.63)**	**1.17 (-0.82)**	1.14 (-0.79)	1.03 (-0.80)	99
		Phi-3	8.44 (-3.68)	13.66 (-4.90)	8.52 (-3.52)	0.57 (-1.11)	0.62 (-0.35)	0.60 (-0.41)	0.59 (-0.41)	149
		XComposer2	9.37 (-12.03)	14.18 (-14.72)	9.92 (-13.04)	0.70 (-2.55)	0.79 (-1.83)	0.75 (-1.79)	0.69 (-1.70)	111
Fr	Fr	LLaVA-NeXT	**13.72 (-8.87)**	**18.78 (-11.53)**	**13.36 (-11.83)**	0.68 (-2.46)	0.71 (-1.59)	0.72 (-1.69)	0.71 (-1.73)	217
		LLaVA-NeXT (FT)	8.56 (-5.24)	12.83 (-5.90)	8.47 (-6.10)	0.67 (-3.01)	0.64 (-2.91)	0.65 (-2.85)	0.64 (-2.78)	91
		Qwen-VL	11.90 (-7.84)	17.07 (-9.49)	11.45 (-9.34)	**0.90 (-1.24)**	**0.76 (-1.42)**	**0.75 (-1.44)**	**0.75 (-1.24)**	170
		Qwen-VL (FT)	9.21 (-10.02)	13.55 (-11.76)	8.83 (-10.33)	0.50 (-5.20)	0.64 (-4.29)	0.64 (-4.24)	0.63 (-4.15)	300
		mPLUG-Owl2	10.58 (-8.18)	15.42 (-10.63)	10.48 (-10.43)	0.58 (-2.17)	0.54 (-1.45)	0.53 (-1.40)	0.52 (-1.31)	142
		Phi-3	8.31 (-3.81)	13.21 (-5.35)	8.06 (-3.99)	0.67 (-1.01)	0.55 (-0.42)	0.52 (-0.48)	0.51 (-0.49)	220
		XComposer2	6.72 (-14.68)	10.82 (-18.08)	6.50 (-16.46)	0.65 (-2.61)	0.50 (-2.12)	0.48 (-2.06)	0.47 (-1.92)	107
En	Ja	LLaVA-NeXT	**9.06 (-13.53)**	**9.11 (-21.19)**	**12.33 (-12.86)**	**0.55 (-2.60)**	**0.55 (-1.75)**	**0.55 (-1.85)**	**0.55 (-1.89)**	212
		LLaVA-NeXT (FT)	0.31 (-13.49)	0.67 (-18.07)	0.41 (-14.16)	0.01 (-3.67)	0.01 (-3.54)	0.01 (-3.49)	0.01 (-3.41)	75
		Qwen-VL	3.53 (-16.21)	3.61 (-22.95)	4.96 (-15.83)	0.18 (-1.96)	0.18 (-2.00)	0.18 (-2.02)	0.18 (-1.81)	132
		Qwen-VL (FT)	0.18 (-19.05)	0.54 (-24.77)	0.19 (-18.96)	0.00 (-5.70)	0.00 (-4.93)	0.00 (-4.88)	0.00 (-4.78)	184
		mPLUG-Owl2	4.34 (-14.42)	4.41 (-21.63)	5.79 (-15.12)	0.26 (-2.48)	0.26 (-1.73)	0.26 (-1.67)	0.26 (-1.57)	106
		Phi-3	2.08 (-10.04)	2.12 (-16.43)	2.74 (-9.31)	0.04 (-1.64)	0.04 (-0.93)	0.04 (-0.96)	0.04 (-0.96)	189
		XComposer2	4.13 (-17.26)	4.18 (-24.72)	5.45 (-17.51)	0.19 (-3.07)	0.19 (-2.43)	0.19 (-2.35)	0.19 (-2.20)	117
Ja	En	LLaVA-NeXT	**12.71 (-9.87)**	**20.20 (-10.11)**	**13.92 (-11.27)**	0.79 (-2.35)	0.88 (-1.42)	0.91 (-1.49)	0.84 (-1.60)	213
		LLaVA-NeXT (FT)	6.14 (-7.67)	10.85 (-7.88)	6.28 (-8.30)	0.43 (-3.25)	0.28 (-3.27)	0.24 (-3.25)	0.22 (-3.20)	82
		Qwen-VL	11.21 (-8.54)	17.07 (-9.49)	11.94 (-8.85)	**0.82 (-1.32)**	**0.92 (-1.26)**	**0.97 (-1.23)**	**0.95 (-1.05)**	114
		Qwen-VL (FT)	5.06 (-14.17)	8.68 (-16.63)	5.08 (-14.07)	0.02 (-5.68)	0.01 (-4.92)	0.03 (-4.85)	0.03 (-4.75)	211
		mPLUG-Owl2	9.20 (-9.56)	15.13 (-10.92)	9.13 (-11.77)	0.65 (-2.10)	0.57 (-1.42)	0.57 (-1.37)	0.56 (-1.27)	82
		Phi-3	5.71 (-6.41)	10.40 (-8.15)	5.75 (-6.30)	0.43 (-1.25)	0.40 (-0.57)	0.48 (-0.52)	0.48 (-0.52)	170
		XComposer2	10.05 (-11.35)	16.41 (-12.49)	10.88 (-12.08)	0.62 (-2.64)	0.85 (-1.77)	0.84 (-1.70)	0.73 (-1.66)	151
Ja	Ja	LLaVA-NeXT	**20.68 (-1.91)**	**20.68 (-9.63)**	**24.14 (-1.05)**	1.04 (-2.11)	1.06 (-1.23)	1.06 (-1.34)	1.06 (-1.37)	247
		LLaVA-NeXT (FT)	14.11 (+0.31)	14.11 (-4.62)	16.72 (+2.15)	1.07 (-2.61)	1.07 (-2.48)	1.07 (-2.43)	1.07 (-2.35)	177
		Qwen-VL	17.32 (-2.43)	17.32 (-9.25)	20.87 (+0.08)	**2.37 (+0.23)**	**2.37 (+0.19)**	**2.37 (+0.17)**	**2.37 (+0.37)**	186
		Qwen-VL (FT)	3.32 (-15.91)	3.32 (-21.99)	3.39 (-15.77)	0.87 (-4.84)	0.87 (-4.06)	0.87 (-4.01)	0.87 (-3.91)	293
		mPLUG-Owl2	11.63 (-7.14)	11.63 (-14.42)	13.86 (-7.04)	1.32 (-1.42)	1.32 (-0.66)	1.32 (-0.61)	1.32 (-0.51)	146
		Phi-3	16.32 (+4.19)	16.32 (-2.24)	18.38 (+6.33)	2.29 (+0.61)	2.29 (+1.32)	2.29 (+1.28)	2.29 (+1.28)	211
		XComposer2	13.43 (-7.97)	13.43 (-15.48)	15.62 (-7.34)	1.48 (-1.78)	1.48 (-1.15)	1.48 (-1.06)	1.48 (-0.91)	167
En	It	LLaVA-NeXT	4.10 (-18.48)	7.54 (-22.76)	5.19 (-20.00)	0.11 (-3.04)	0.07 (-2.23)	0.06 (-2.34)	0.09 (-2.35)	118
		LLaVA-NeXT (FT)	6.14 (-7.66)	9.31 (-9.42)	7.47 (-7.10)	**0.79 (-2.89)**	**0.86 (-2.68)**	**0.86 (-2.63)**	**0.86 (-2.56)**	72
		Qwen-VL	**7.33 (-12.42)**	**12.36 (-14.21)**	**8.45 (-12.34)**	0.33 (-1.81)	0.35 (-1.83)	0.38 (-1.82)	0.43 (-1.57)	116
		Qwen-VL (FT)	5.59 (-13.64)	8.95 (-16.36)	7.13 (-12.03)	0.48 (-5.22)	0.50 (-4.43)	0.51 (-4.38)	0.50 (-4.28)	176
		mPLUG-Owl2	5.53 (-13.23)	10.27 (-15.77)	6.02 (-14.88)	0.18 (-2.56)	0.18 (-1.80)	0.19 (-1.74)	0.32 (-1.51)	103
		Phi-3	4.50 (-7.62)	8.41 (-10.15)	4.92 (-7.13)	0.13 (-1.55)	0.11 (-0.86)	0.12 (-0.89)	0.14 (-0.86)	171
		XComposer2	4.42 (-16.98)	8.51 (-20.39)	4.85 (-18.11)	0.12 (-3.13)	0.11 (-2.51)	0.11 (-2.43)	0.19 (-2.20)	77
It	En	LLaVA-NeXT	9.95 (-12.64)	17.66 (-12.65)	10.50 (-14.69)	0.66 (-2.49)	0.42 (-1.87)	0.39 (-2.01)	0.38 (-2.05)	219
		LLaVA-NeXT (FT)	10.00 (-3.81)	15.32 (-3.41)	10.04 (-4.53)	**1.36 (-2.33)**	**1.34 (-2.21)**	**1.28 (-2.21)**	**1.26 (-2.16)**	92
		Qwen-VL	**12.49 (-7.25)**	**18.79 (-7.78)**	**13.38 (-7.40)**	0.90 (-1.24)	1.04 (-1.14)	1.02 (-1.17)	0.99 (-1.00)	93
		Qwen-VL (FT)	2.31 (-16.92)	6.22 (-19.09)	3.01 (-16.15)	0.16 (-5.54)	0.07 (-4.86)	0.06 (-4.82)	0.05 (-4.73)	149
		mPLUG-Owl2	9.42 (-9.35)	15.46 (-10.58)	9.71 (-11.20)	0.70 (-2.05)	0.57 (-1.42)	0.52 (-1.42)	0.47 (-1.36)	56
		Phi-3	6.05 (-6.07)	11.51 (-7.04)	6.13 (-5.92)	0.48 (-1.20)	0.51 (-0.46)	0.53 (-0.47)	0.54 (-0.46)	141
		XComposer2	8.43 (-12.97)	14.12 (-14.78)	8.57 (-14.39)	0.49 (-2.76)	0.58 (-2.05)	0.56 (-1.98)	0.52 (-1.87)	88
It	It	LLaVA-NeXT	**11.91 (-10.67)**	**17.01 (-13.29)**	9.89 (-15.30)	**2.69 (-0.45)**	**1.84 (-0.46)**	**1.82 (-0.58)**	**1.81 (-0.63)**	196
		LLaVA-NeXT (FT)	7.64 (-6.16)	11.51 (-7.22)	5.84 (-8.73)	1.94 (-1.74)	1.15 (-2.40)	1.10 (-2.39)	1.09 (-2.33)	93
		Qwen-VL	9.27 (-10.47)	13.87 (-12.70)	**10.28 (-10.51)**	1.71 (-0.43)	0.92 (-1.26)	0.90 (-1.29)	0.88 (-1.11)	116
		Qwen-VL (FT)	3.43 (-15.80)	6.54 (-18.77)	3.71 (-15.44)	0.03 (-5.67)	0.06 (-4.87)	0.06 (-4.82)	0.06 (-4.72)	299
		mPLUG-Owl2	5.81 (-12.95)	9.44 (-16.60)	5.93 (-14.97)	0.65 (-2.09)	0.36 (-1.63)	0.41 (-1.52)	0.39 (-1.44)	102
		Phi-3	6.30 (-5.82)	10.20 (-8.35)	5.60 (-6.44)	0.46 (-1.22)	0.24 (-0.73)	0.22 (-0.78)	0.22 (-0.78)	226
		XComposer2	4.35 (-17.05)	6.87 (-22.03)	4.30 (-18.66)	0.71 (-2.55)	0.32 (-2.31)	0.31 (-2.23)	0.31 (-2.08)	66

Table 9: Results of LVLMs in Alignment-5 Task. Bold fonts indicate the best score for that language combination. The values are noted next to the output of the difference by the same model in the method with instruction and output in English ({En}-{En}). Red indicates a higher value than that method; blue indicates a lower value.

Input	Output	LVLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
			exact	partial		n=0	n=1	n=2	n=∞	
En	En	Phi-3	9.74	13.75	8.06	0.95	0.86	0.92	0.90	108
		XComposer2	**16.57**	**21.56**	**16.53**	**1.51**	**1.47**	**1.43**	**1.36**	223
En	Es	Phi-3	**10.35 (+0.61)**	**14.46 (+0.72)**	**11.67 (+3.61)**	**1.20 (+0.25)**	**1.37 (+0.51)**	**1.43 (+0.51)**	**1.38 (+0.48)**	141
		XComposer2	10.03 (-6.55)	13.75 (-7.81)	10.04 (-6.49)	0.79 (-0.73)	0.96 (-0.51)	0.92 (-0.50)	0.89 (-0.47)	116
Es	En	Phi-3	7.99 (-1.75)	14.53 (+0.78)	7.35 (-0.71)	**0.90 (-0.05)**	**0.79 (-0.07)**	**0.80 (-0.12)**	**0.75 (-0.16)**	91
		XComposer2	**9.58 (-6.99)**	**15.05 (-6.51)**	**9.03 (-7.50)**	0.67 (-0.85)	0.55 (-0.92)	0.61 (-0.82)	0.51 (-0.85)	69
Es	Es	Phi-3	**12.81 (+3.07)**	**16.25 (+2.50)**	**13.04 (+4.98)**	1.10 (+0.16)	1.23 (+0.37)	**1.24 (+0.32)**	**1.21 (+0.31)**	190
		XComposer2	9.19 (-7.38)	11.93 (-9.63)	8.68 (-7.85)	**1.11 (-0.40)**	**1.24 (-0.24)**	1.17 (-0.26)	1.14 (-0.21)	77
En	Fr	Phi-3	9.00 (-0.74)	14.17 (+0.42)	**9.84 (+1.77)**	**1.60 (+0.65)**	**1.12 (+0.26)**	**1.11 (+0.19)**	**1.03 (+0.13)**	151
		XComposer2	**9.32 (-7.25)**	**14.32 (-7.24)**	9.28 (-7.25)	1.09 (-0.42)	0.89 (-0.59)	0.81 (-0.62)	0.74 (-0.62)	79
Fr	En	Phi-3	10.37 (+0.63)	16.77 (+3.03)	9.83 (+1.77)	**1.22 (+0.27)**	**1.13 (+0.26)**	**1.10 (+0.17)**	**1.11 (+0.21)**	154
		XComposer2	**10.79 (-5.78)**	**17.10 (-4.46)**	**10.12 (-6.41)**	0.72 (-0.79)	0.64 (-0.83)	0.62 (-0.81)	0.63 (-0.73)	121
Fr	Fr	Phi-3	**11.06 (+1.32)**	**16.04 (+2.30)**	**10.21 (+2.15)**	**0.95 (+0.00)**	**0.87 (+0.01)**	**0.84 (-0.08)**	**0.86 (-0.04)**	220
		XComposer2	9.90 (-6.67)	14.07 (-7.49)	8.93 (-7.60)	0.73 (-0.79)	0.61 (-0.86)	0.56 (-0.87)	0.50 (-0.85)	116
En	De	Phi-3	**9.55 (-0.19)**	**13.14 (-0.61)**	**11.88 (+3.82)**	**0.80 (-0.15)**	**0.80 (-0.07)**	**0.80 (-0.13)**	**0.95 (+0.04)**	216
		XComposer2	8.02 (-8.55)	11.26 (-10.30)	9.48 (-7.05)	0.78 (-0.73)	0.77 (-0.70)	0.75 (-0.68)	0.78 (-0.57)	107
De	En	Phi-3	9.68 (-0.06)	16.53 (+2.78)	**10.01 (+1.95)**	0.67 (-0.27)	0.62 (-0.24)	0.70 (-0.22)	0.72 (-0.19)	183
		XComposer2	**11.05 (-5.52)**	**17.07 (-4.49)**	9.39 (-7.13)	**0.95 (-0.56)**	**0.89 (-0.58)**	**0.90 (-0.52)**	**0.91 (-0.45)**	86
De	De	Phi-3	**13.37 (+3.63)**	**16.55 (+2.80)**	**15.57 (+7.51)**	**1.72 (+0.77)**	**1.50 (+0.64)**	**1.53 (+0.61)**	**1.55 (+0.65)**	240
		XComposer2	9.89 (-6.68)	11.93 (-9.63)	11.73 (-4.80)	1.17 (-0.34)	0.91 (-0.56)	0.81 (-0.62)	0.80 (-0.56)	107
En	It	Phi-3	**4.98 (-4.76)**	**8.61 (-5.14)**	**5.63 (-2.43)**	0.08 (-0.87)	0.12 (-0.74)	0.14 (-0.78)	0.12 (-0.78)	150
		XComposer2	4.58 (-11.99)	8.40 (-13.16)	5.44 (-11.09)	**0.35 (-1.16)**	**0.19 (-1.28)**	**0.19 (-1.23)**	**0.19 (-1.16)**	87
It	En	Phi-3	7.11 (-2.63)	13.70 (-0.05)	7.04 (-1.03)	**0.89 (-0.06)**	**0.95 (+0.08)**	**0.95 (+0.03)**	**0.93 (+0.03)**	143
		XComposer2	**9.82 (-6.75)**	**16.56 (-5.00)**	**8.72 (-7.81)**	0.58 (-0.93)	0.64 (-0.83)	0.68 (-0.75)	0.64 (-0.71)	94
It	It	Phi-3	**7.93 (-1.81)**	**11.61 (-2.14)**	**9.17 (+1.11)**	0.14 (-0.81)	0.11 (-0.75)	0.11 (-0.81)	0.11 (-0.79)	183
		XComposer2	4.29 (-12.29)	6.92 (-14.64)	5.27 (-11.26)	**0.22 (-1.29)**	**0.18 (-1.29)**	**0.18 (-1.25)**	**0.18 (-1.18)**	65
En	Nl	Phi-3	4.07 (-5.67)	8.41 (-5.34)	5.03 (-3.03)	0.04 (-0.91)	0.04 (-0.82)	0.03 (-0.89)	0.05 (-0.85)	240
		XComposer2	**5.56 (-11.01)**	**10.07 (-11.49)**	**5.84 (-10.69)**	**0.29 (-1.22)**	**0.52 (-0.95)**	**0.50 (-0.92)**	**0.52 (-0.84)**	78
Nl	En	Phi-3	4.09 (-5.65)	9.41 (-4.34)	4.17 (-3.89)	0.50 (-0.45)	0.35 (-0.51)	0.32 (-0.61)	0.31 (-0.59)	213
		XComposer2	**9.74 (-6.83)**	**17.17 (-4.39)**	**9.44 (-7.09)**	**1.21 (-0.30)**	**0.99 (-0.48)**	**0.98 (-0.45)**	**0.93 (-0.43)**	92
Nl	Nl	Phi-3	8.37 (-1.37)	13.12 (-0.63)	6.26 (-1.81)	0.01 (-0.94)	**1.38 (+0.52)**	**1.38 (+0.46)**	**1.38 (+0.47)**	273
		XComposer2	**10.73 (-5.85)**	**14.59 (-6.97)**	**9.95 (-6.58)**	**0.08 (-1.44)**	0.92 (-0.55)	0.94 (-0.49)	0.94 (-0.42)	73
En	Sv	Phi-3	6.55 (-3.19)	9.20 (-4.55)	6.18 (-1.88)	0.05 (-0.89)	0.04 (-0.82)	0.04 (-0.88)	0.03 (-0.87)	235
		XComposer2	**8.03 (-8.55)**	**10.90 (-10.66)**	**6.91 (-9.62)**	**0.31 (-1.20)**	**0.32 (-1.16)**	**0.30 (-1.12)**	**0.29 (-1.07)**	76
Sv	En	Phi-3	4.69 (-5.05)	10.04 (-3.71)	4.58 (-3.48)	0.51 (-0.44)	0.46 (-0.41)	0.51 (-0.41)	0.45 (-0.45)	176
		XComposer2	**11.35 (-5.23)**	**13.91 (-7.65)**	**8.60 (-7.93)**	**1.23 (-0.28)**	**0.48 (-0.99)**	**0.56 (-0.87)**	**0.52 (-0.83)**	78
Sv	Sv	Phi-3	**14.03 (+4.29)**	**15.53 (+1.78)**	**9.92 (+1.85)**	0.74 (-0.21)	0.27 (-0.60)	0.26 (-0.66)	0.26 (-0.65)	194
		XComposer2	11.58 (-4.99)	13.07 (-8.50)	8.41 (-8.12)	**1.26 (-0.26)**	**0.44 (-1.03)**	**0.44 (-0.99)**	**0.43 (-0.93)**	63
En	Ru	Phi-3	0.61 (-9.13)	2.17 (-11.58)	0.31 (-7.75)	**0.00 (-0.95)**	0.00 (-0.86)	0.00 (-0.92)	0.00 (-0.90)	194
		XComposer2	**3.70 (-12.88)**	**6.50 (-15.06)**	**3.07 (-13.46)**	**0.00 (-1.51)**	**0.00 (-1.47)**	**0.01 (-1.42)**	**0.01 (-1.34)**	73
Ru	En	Phi-3	**6.62 (-3.12)**	**12.82 (-0.92)**	**6.57 (-1.50)**	**0.31 (-0.63)**	**0.47 (-0.39)**	**0.47 (-0.45)**	**0.46 (-0.44)**	147
		XComposer2	4.69 (-11.89)	8.56 (-13.00)	4.07 (-12.46)	0.19 (-1.32)	0.11 (-1.37)	0.14 (-1.28)	0.10 (-1.25)	62
Ru	Ru	Phi-3	2.42 (-7.32)	4.45 (-9.30)	1.58 (-6.48)	**0.21 (-0.74)**	**0.21 (-0.66)**	**0.20 (-0.72)**	**0.20 (-0.70)**	269
		XComposer2	**2.73 (-13.84)**	**4.85 (-16.71)**	**2.22 (-14.31)**	0.17 (-1.35)	0.16 (-1.31)	0.15 (-1.28)	0.14 (-1.21)	45
En	Ja	Phi-3	2.53 (-7.21)	2.53 (-11.22)	3.26 (-4.80)	0.06 (-0.89)	0.06 (-0.80)	0.06 (-0.86)	0.06 (-0.84)	202
		XComposer2	**3.27 (-13.31)**	**3.27 (-18.29)**	**3.78 (-12.75)**	**0.21 (-1.30)**	**0.21 (-1.26)**	**0.21 (-1.22)**	**0.21 (-1.15)**	109
Ja	En	Phi-3	8.17 (-1.57)	15.38 (+1.64)	7.46 (-0.60)	0.37 (-0.58)	0.40 (-0.46)	0.44 (-0.48)	0.38 (-0.52)	168
		XComposer2	**10.59 (-5.99)**	**18.39 (-3.17)**	**9.71 (-6.82)**	**0.61 (-0.90)**	**0.57 (-0.90)**	**0.54 (-0.89)**	**0.50 (-0.86)**	159
Ja	Ja	Phi-3	**8.73 (-1.01)**	**8.74 (-5.01)**	**12.19 (+4.13)**	0.88 (-0.06)	0.88 (+0.02)	0.88 (-0.04)	0.88 (-0.02)	214
		XComposer2	6.04 (-10.53)	6.04 (-15.52)	7.18 (-9.34)	**1.22 (-0.29)**	**1.22 (-0.25)**	**1.22 (-0.20)**	**1.22 (-0.13)**	133
En	Zh	Phi-3	4.48 (-5.26)	4.52 (-9.23)	5.14 (-2.92)	0.13 (-0.81)	0.14 (-0.73)	0.14 (-0.79)	0.14 (-0.77)	145
		XComposer2	**13.35 (-3.23)**	**13.38 (-8.18)**	**16.00 (-0.53)**	**0.69 (-0.83)**	**0.68 (-0.80)**	**0.67 (-0.75)**	**0.67 (-0.69)**	124
Zh	En	Phi-3	3.74 (-6.00)	3.74 (-10.01)	6.11 (-1.95)	0.14 (-0.81)	0.14 (-0.72)	0.14 (-0.78)	0.14 (-0.76)	186
		XComposer2	**9.27 (-7.30)**	**9.27 (-12.29)**	**11.95 (-4.58)**	**0.25 (-1.26)**	**0.24 (-1.23)**	**0.24 (-1.18)**	**0.24 (-1.11)**	215
Zh	Zh	Phi-3	2.53 (-7.21)	2.53 (-11.22)	**4.94 (-3.12)**	0.00 (-0.95)	0.00 (-0.86)	0.00 (-0.92)	0.00 (-0.90)	55
		XComposer2	**2.87 (-13.70)**	**2.87 (-18.69)**	4.44 (-12.09)	**0.09 (-1.42)**	**0.09 (-1.38)**	**0.09 (-1.34)**	**0.09 (-1.26)**	55

Table 10: Results for Phi-3 and XComposer2 in the Alignment-10 task. Bold fonts indicate the best score for that language combination. The values are noted next to the output of the difference by the same model in the method with instruction and output in English ({En}-{En}). Red indicates a higher value than that method; blue indicates a lower value.

Input	Output	LVLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
			exact	partial		n=0	n=1	n=2	n=∞	
En	En	LLaVA-NeXT	**26.49**	**31.54**	**26.07**	1.35	1.65	1.66	1.70	252
		LLaVA-NeXT (FT)	16.98	22.70	19.95	3.02	3.33	3.23	3.11	83
		Qwen-VL	21.11	27.03	21.78	1.60	1.59	1.56	1.52	155
		Qwen-VL (FT)	21.12	24.87	21.95	**3.57**	**3.83**	**3.78**	**3.68**	177
		mPLUG-Owl2	12.79	17.08	13.48	2.07	1.68	1.59	1.56	151
Es	En	LLaVA-NeXT	4.22 (-22.26)	6.22 (-25.33)	4.17 (-21.90)	0.00 (-1.34)	0.00 (-1.65)	0.00 (-1.66)	0.00 (-1.70)	180
		LLaVA-NeXT (FT)	13.64 (-3.34)	18.60 (-4.10)	13.54 (-6.41)	1.35 (-1.67)	1.44 (-1.89)	1.35 (-1.88)	1.35 (-1.76)	123
		Qwen-VL	10.38 (-10.73)	16.82 (-10.21)	10.40 (-11.38)	0.72 (-0.88)	0.62 (-0.96)	0.64 (-0.92)	0.55 (-0.97)	60
		Qwen-VL (FT)	**15.50 (-5.62)**	**22.52 (-2.34)**	**16.19 (-5.76)**	1.31 (-2.27)	**1.70 (-2.13)**	**1.76 (-2.01)**	**1.65 (-2.02)**	199
		mPLUG-Owl2	11.97 (-0.82)	18.26 (+1.18)	11.68 (-1.80)	**1.72 (-0.35)**	1.28 (-0.40)	1.17 (-0.42)	1.11 (-0.45)	54
Fr	En	LLaVA-NeXT	**21.11 (-5.38)**	**28.40 (-3.14)**	**21.22 (-4.85)**	**2.01 (+0.67)**	**2.04 (+0.38)**	**2.09 (+0.43)**	**1.97 (+0.27)**	232
		LLaVA-NeXT (FT)	15.34 (-1.64)	21.24 (-1.46)	15.24 (-4.71)	1.21 (-1.81)	1.11 (-2.22)	1.08 (-2.15)	1.00 (-2.11)	101
		Qwen-VL	14.24 (-6.86)	20.10 (-6.93)	15.24 (-6.54)	1.43 (-0.17)	1.38 (-0.21)	1.32 (-0.24)	1.27 (-0.25)	144
		Qwen-VL (FT)	12.92 (-8.19)	21.49 (-3.38)	11.85 (-10.10)	0.35 (-3.22)	0.66 (-3.17)	0.64 (-3.14)	0.62 (-3.05)	275
		mPLUG-Owl2	15.56 (+2.77)	22.04 (+4.96)	14.34 (+0.86)	1.69 (-0.38)	1.61 (-0.06)	1.65 (+0.06)	1.47 (-0.09)	106
De	En	LLaVA-NeXT	**21.88 (-4.60)**	**30.48 (-1.07)**	**21.76 (-4.31)**	0.89 (-0.46)	**1.33 (-0.33)**	**1.53 (-0.13)**	1.50 (-0.20)	239
		LLaVA-NeXT (FT)	14.15 (-2.83)	19.99 (-2.71)	12.43 (-7.52)	0.68 (-2.33)	0.90 (-2.43)	0.86 (-2.37)	0.85 (-2.26)	128
		Qwen-VL	17.36 (-3.74)	25.21 (-1.82)	16.82 (-4.95)	**1.11 (-0.49)**	1.31 (-0.28)	1.49 (-0.07)	**1.51 (-0.01)**	109
		Qwen-VL (FT)	13.60 (-7.51)	21.30 (-3.57)	13.17 (-8.77)	0.80 (-2.77)	1.12 (-2.72)	1.14 (-2.64)	1.07 (-2.61)	265
		mPLUG-Owl2	13.29 (+0.50)	19.74 (+2.66)	12.42 (-1.06)	0.78 (-1.28)	0.77 (-0.91)	0.80 (-0.79)	0.77 (-0.79)	75
It	En	LLaVA-NeXT	7.98 (-18.51)	11.14 (-20.41)	5.40 (-20.67)	0.22 (-1.12)	0.28 (-1.38)	0.28 (-1.38)	0.28 (-1.42)	137
		LLaVA-NeXT (FT)	**12.42 (-4.56)**	18.65 (-4.05)	**11.97 (-7.98)**	1.15 (-1.86)	1.11 (-2.23)	1.07 (-2.16)	1.00 (-2.10)	105
		Qwen-VL	11.02 (-10.09)	**18.70 (-8.32)**	10.89 (-10.89)	**1.18 (-0.42)**	**1.17 (-0.42)**	**1.14 (-0.41)**	**1.10 (-0.42)**	100
		Qwen-VL (FT)	0.00 (-21.12)	0.00 (-24.87)	0.00 (-21.95)	0.00 (-3.57)	0.00 (-3.83)	0.00 (-3.78)	0.00 (-3.68)	83
		mPLUG-Owl2	10.03 (-2.75)	17.28 (+0.20)	9.77 (-3.71)	1.09 (-0.98)	0.90 (-0.78)	0.84 (-0.74)	0.82 (-0.74)	55
Nl	En	LLaVA-NeXT	**15.81 (-10.68)**	**24.80 (-6.75)**	**21.13 (-4.94)**	0.11 (-1.24)	**2.04 (+0.38)**	**2.29 (+0.63)**	**1.83 (+0.13)**	223
		LLaVA-NeXT (FT)	9.92 (-7.06)	15.45 (-7.25)	9.73 (-10.22)	0.91 (-2.11)	0.98 (-2.36)	0.91 (-2.32)	0.88 (-2.23)	153
		Qwen-VL	11.65 (-9.46)	18.87 (-8.16)	13.19 (-8.59)	1.54 (-0.06)	1.47 (-0.12)	1.48 (-0.08)	1.44 (-0.08)	136
		Qwen-VL (FT)	10.35 (-10.76)	16.35 (-8.52)	10.79 (-11.16)	0.70 (-2.87)	1.13 (-2.70)	1.07 (-2.71)	1.04 (-2.64)	331
		mPLUG-Owl2	12.19 (-0.59)	20.44 (+3.36)	12.97 (-0.51)	**1.81 (-0.25)**	1.61 (-0.06)	1.56 (-0.03)	1.50 (-0.06)	82
Sv	En	LLaVA-NeXT	**18.70 (-7.79)**	**25.48 (-6.07)**	**18.98 (-7.09)**	1.79 (+0.44)	**1.86 (+0.20)**	**1.86 (+0.20)**	**1.80 (+0.10)**	246
		LLaVA-NeXT (FT)	9.30 (-7.68)	14.68 (-8.02)	9.12 (-10.83)	0.80 (-2.22)	0.76 (-2.57)	0.73 (-2.50)	0.71 (-2.40)	141
		Qwen-VL	11.77 (-9.33)	17.73 (-9.30)	13.03 (-8.75)	1.57 (-0.02)	1.36 (-0.23)	1.30 (-0.26)	1.25 (-0.27)	107
		Qwen-VL (FT)	11.00 (-10.11)	17.97 (-6.89)	9.83 (-12.12)	0.69 (-2.88)	0.74 (-3.10)	0.73 (-3.05)	0.63 (-3.05)	233
		mPLUG-Owl2	12.49 (-0.29)	18.51 (+1.43)	11.40 (-2.08)	**1.90 (-0.17)**	1.29 (-0.39)	1.29 (-0.30)	1.21 (-0.35)	81
Ru	En	LLaVA-NeXT	**18.31 (-8.18)**	**26.30 (-5.25)**	**18.43 (-7.64)**	**1.68 (+0.34)**	**1.64 (-0.01)**	**1.65 (-0.01)**	**1.59 (-0.11)**	241
		LLaVA-NeXT (FT)	9.61 (-7.37)	13.42 (-9.28)	8.54 (-11.41)	1.01 (-2.01)	0.97 (-2.36)	0.94 (-2.29)	0.91 (-2.20)	125
		Qwen-VL	13.36 (-7.75)	20.75 (-6.28)	13.97 (-7.81)	1.02 (-0.57)	1.13 (-0.45)	1.25 (-0.31)	1.22 (-0.30)	128
		Qwen-VL (FT)	9.66 (-11.45)	15.91 (-8.96)	9.12 (-12.83)	0.90 (-2.67)	0.87 (-2.96)	0.98 (-2.80)	0.87 (-2.81)	258
		mPLUG-Owl2	12.56 (-0.22)	19.45 (+2.37)	12.60 (-0.88)	1.60 (-0.47)	1.60 (-0.07)	1.51 (-0.07)	1.41 (-0.15)	96
Ja	En	LLaVA-NeXT	**15.36 (-11.13)**	**24.41 (-7.13)**	**16.18 (-9.89)**	**1.12 (-0.23)**	**1.15 (-0.51)**	**1.28 (-0.38)**	**1.11 (-0.59)**	208
		LLaVA-NeXT (FT)	7.69 (-9.28)	12.61 (-10.09)	8.29 (-11.66)	0.85 (-2.17)	0.54 (-2.80)	0.47 (-2.76)	0.45 (-2.65)	68
		Qwen-VL	10.32 (-10.78)	17.64 (-9.38)	9.75 (-12.03)	0.97 (-0.63)	0.75 (-0.83)	0.78 (-0.78)	0.76 (-0.76)	108
		Qwen-VL (FT)	0.73 (-20.38)	3.14 (-21.72)	0.00 (-21.95)	0.00 (-3.57)	0.00 (-3.83)	0.00 (-3.78)	0.00 (-3.68)	153
		mPLUG-Owl2	10.02 (-2.76)	17.10 (+0.02)	8.27 (-5.21)	1.11 (-0.96)	0.70 (-0.98)	0.71 (-0.88)	0.67 (-0.89)	76
Zh	En	LLaVA-NeXT	**13.44 (-2.42)**	**21.98 (+1.04)**	**12.83 (-3.67)**	**0.96 (-0.12)**	**1.43 (+0.19)**	**1.58 (+0.37)**	**1.42 (+0.17)**	168
		LLaVA-NeXT (FT)	6.71 (-4.93)	13.24 (-2.23)	6.44 (-7.69)	0.88 (-1.54)	0.59 (-2.01)	0.57 (-1.94)	0.54 (-1.89)	94
		Qwen-VL	8.98 (-3.15)	16.16 (-0.26)	9.65 (-2.54)	0.48 (-0.85)	0.45 (-0.72)	0.46 (-0.66)	0.40 (-0.67)	138
		Qwen-VL (FT)	8.90 (-4.87)	16.99 (-0.64)	8.79 (-4.64)	0.14 (-2.84)	0.10 (-3.00)	0.10 (-2.96)	0.09 (-2.91)	242
		mPLUG-Owl2	5.25 (-7.57)	11.81 (-5.33)	4.32 (-9.20)	0.15 (-1.93)	0.14 (-1.54)	0.15 (-1.45)	0.15 (-1.42)	34

Table 11: Results of LVLMs in Alignment-10 Task, which the method is an instruction in languages other than English and output in English ({Lang}-{En}). Bold fonts indicate the best score for that language combination. The values are noted next to the output of the difference by the same model in the method with instruction and output in English ({En}-{En}). Red indicates a higher value than that method; blue indicates a lower value.

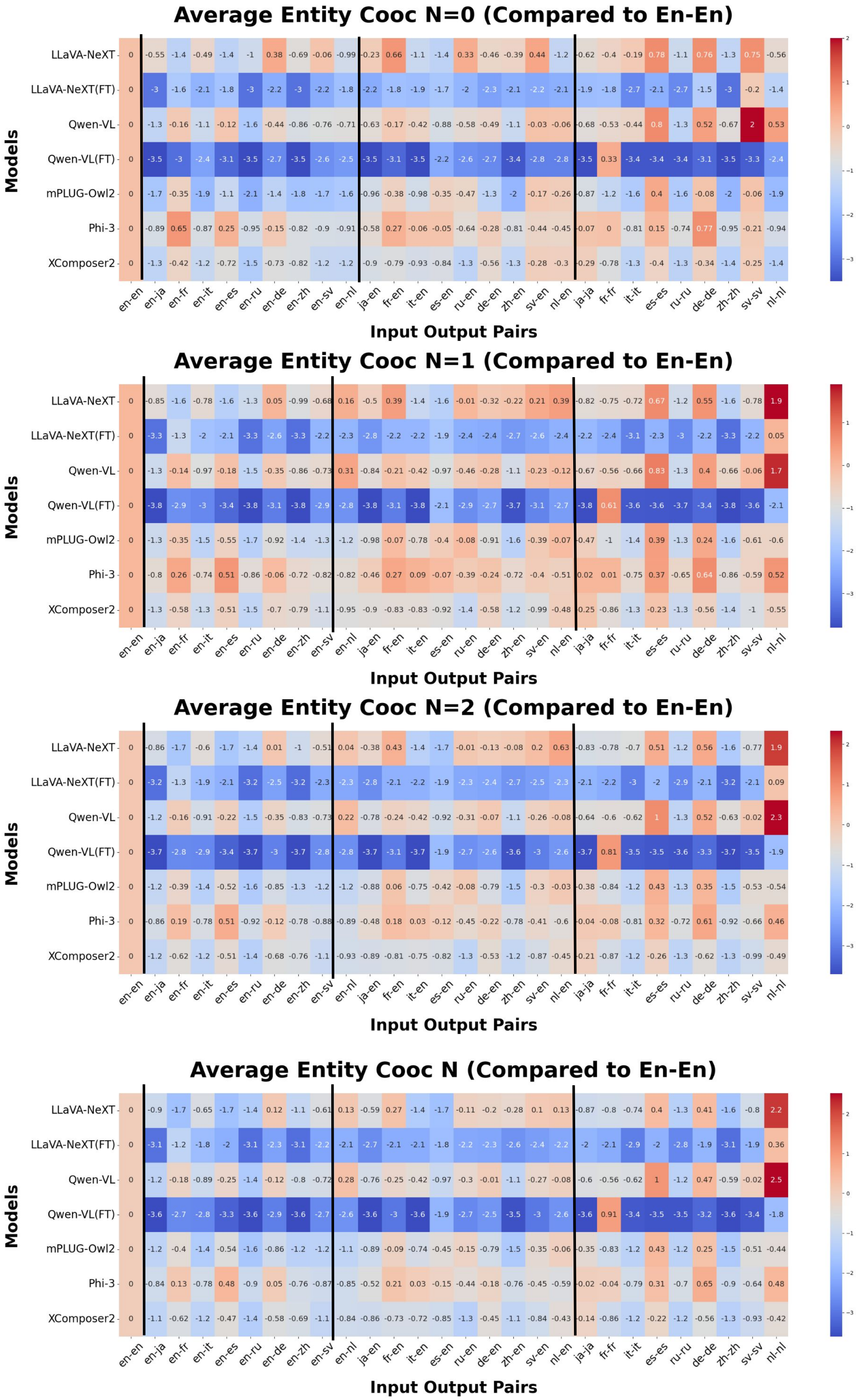

Figure 5: Visualization of Alignment-10 results in a heat map. We made the visualization based on when we had LVLMs give instructions and output in English.

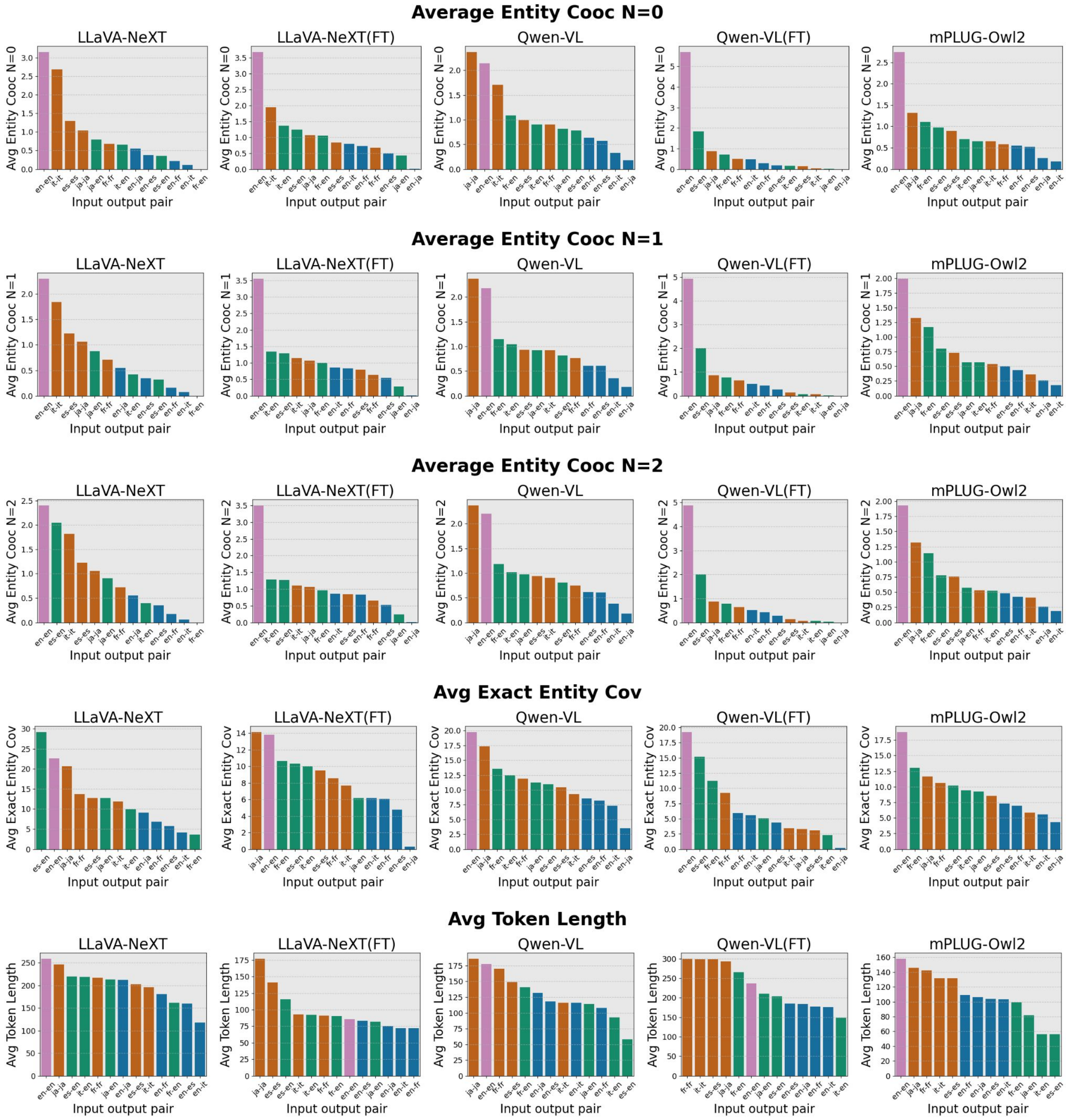

Figure 6: The rest of the results in the Alignment-5 task. From this figure, it can also be seen that the English instructions are optimal, even if the number of data is expanded. Purple bin indicates the method which is the instruction and the output in English ({En}-{En}), Green bin indicates the instruction in languages other than English and the output in English ({Lang}-{En}), Brown bin indicates the instruction and output in languages other than English ({Lang}-{Lang}) and Blue bin indicates the instruction in English and the output in languages other than English ({En}-{Lang}).

Input	Output	LVLM	BLEU	ROUGE			BertScore
				1	2	L	
En	En	LLaVA-NeXT	0.01	0.24	0.05	0.15	0.82
		LLaVA-NeXT (FT)	**0.07**	0.28	**0.13**	**0.22**	**0.85**
		Qwen-VL	0.01	0.22	0.05	0.14	0.82
		Qwen-VL (FT)	0.06	**0.28**	0.12	0.22	0.84
		mPLUG-Owl2	0.01	0.24	0.05	0.15	0.82
		Phi-3	0.01	0.20	0.04	0.12	0.82
		XComposer2	0.01	0.24	0.05	0.14	0.82
En	Es	LLaVA-NeXT	**0.01 (-0.00)**	**0.28 (+0.04)**	**0.06 (+0.01)**	**0.16 (+0.01)**	**0.81 (-0.01)**
		LLaVA-NeXT (FT)	0.00 (-0.07)	0.05 (-0.23)	0.01 (-0.12)	0.04 (-0.18)	0.78 (-0.07)
		Qwen-VL	0.00 (-0.01)	0.20 (-0.03)	0.04 (-0.01)	0.12 (-0.02)	0.80 (-0.02)
		Qwen-VL (FT)	0.00 (-0.06)	0.03 (-0.25)	0.00 (-0.11)	0.03 (-0.19)	0.77 (-0.07)
		mPLUG-Owl2	0.00 (-0.01)	0.22 (-0.03)	0.04 (-0.01)	0.13 (-0.02)	0.80 (-0.02)
		Phi-3	0.00 (-0.00)	0.21 (+0.01)	0.04 (+0.00)	0.13 (+0.00)	0.79 (-0.02)
		XComposer2	0.00 (-0.01)	0.18 (-0.06)	0.04 (-0.02)	0.11 (-0.03)	0.80 (-0.02)
En	Fr	LLaVA-NeXT	**0.00 (-0.01)**	**0.20 (-0.04)**	**0.04 (-0.02)**	**0.12 (-0.03)**	**0.79 (-0.02)**
		LLaVA-NeXT (FT)	0.00 (-0.07)	0.06 (-0.21)	0.02 (-0.11)	0.05 (-0.17)	0.78 (-0.06)
		Qwen-VL	0.00 (-0.01)	0.15 (-0.08)	0.03 (-0.02)	0.09 (-0.05)	0.79 (-0.03)
		Qwen-VL (FT)	0.00 (-0.06)	0.03 (-0.25)	0.00 (-0.11)	0.03 (-0.19)	0.77 (-0.07)
		mPLUG-Owl2	0.00 (-0.01)	0.16 (-0.08)	0.03 (-0.02)	0.10 (-0.05)	0.79 (-0.03)
		Phi-3	0.00 (-0.00)	0.15 (-0.04)	0.02 (-0.01)	0.09 (-0.03)	0.78 (-0.03)
		XComposer2	0.00 (-0.01)	0.03 (-0.21)	0.01 (-0.05)	0.03 (-0.12)	0.78 (-0.04)
En	De	LLaVA-NeXT	**0.00 (-0.01)**	**0.20 (-0.05)**	**0.03 (-0.02)**	**0.11 (-0.03)**	**0.80 (-0.02)**
		LLaVA-NeXT (FT)	0.00 (-0.07)	0.04 (-0.23)	0.01 (-0.12)	0.03 (-0.19)	0.76 (-0.08)
		Qwen-VL	0.00 (-0.01)	0.14 (-0.08)	0.02 (-0.03)	0.09 (-0.06)	0.79 (-0.03)
		Qwen-VL (FT)	0.00 (-0.06)	0.03 (-0.25)	0.00 (-0.11)	0.03 (-0.19)	0.76 (-0.08)
		mPLUG-Owl2	0.00 (-0.01)	0.14 (-0.10)	0.02 (-0.03)	0.09 (-0.07)	0.79 (-0.03)
		Phi-3	0.00 (-0.00)	0.14 (-0.05)	0.02 (-0.02)	0.09 (-0.03)	0.78 (-0.03)
		XComposer2	0.00 (-0.01)	0.14 (-0.10)	0.02 (-0.03)	0.09 (-0.06)	0.79 (-0.03)
En	It	LLaVA-NeXT	**0.00 (-0.01)**	**0.19 (-0.05)**	**0.02 (-0.03)**	**0.11 (-0.04)**	**0.80 (-0.01)**
		LLaVA-NeXT (FT)	0.00 (-0.07)	0.04 (-0.23)	0.01 (-0.12)	0.04 (-0.19)	0.77 (-0.08)
		Qwen-VL	0.00 (-0.01)	0.14 (-0.08)	0.02 (-0.03)	0.09 (-0.06)	0.80 (-0.02)
		Qwen-VL (FT)	0.00 (-0.06)	0.04 (-0.24)	0.01 (-0.11)	0.04 (-0.18)	0.76 (-0.07)
		mPLUG-Owl2	0.00 (-0.01)	0.14 (-0.10)	0.02 (-0.04)	0.09 (-0.07)	0.80 (-0.02)
		Phi-3	0.00 (-0.00)	0.10 (-0.09)	0.01 (-0.03)	0.07 (-0.05)	0.78 (-0.03)
		XComposer2	0.00 (-0.01)	0.10 (-0.14)	0.01 (-0.04)	0.07 (-0.07)	0.80 (-0.02)
En	Nl	LLaVA-NeXT	0.00 (-0.01)	**0.23 (-0.01)**	**0.04 (-0.01)**	**0.15 (-0.00)**	**0.81 (-0.01)**
		LLaVA-NeXT (FT)	**0.01 (-0.06)**	0.12 (-0.15)	0.03 (-0.10)	0.09 (-0.13)	0.78 (-0.07)
		Qwen-VL	0.00 (-0.01)	0.20 (-0.03)	0.04 (-0.01)	0.13 (-0.01)	0.80 (-0.02)
		Qwen-VL (FT)	0.00 (-0.06)	0.06 (-0.23)	0.01 (-0.11)	0.05 (-0.17)	0.76 (-0.08)
		mPLUG-Owl2	0.00 (-0.01)	0.17 (-0.07)	0.03 (-0.02)	0.11 (-0.04)	0.80 (-0.03)
		Phi-3	0.00 (-0.00)	0.10 (-0.10)	0.01 (-0.02)	0.08 (-0.05)	0.77 (-0.05)
		XComposer2	0.00 (-0.01)	0.15 (-0.09)	0.03 (-0.03)	0.11 (-0.04)	0.80 (-0.02)
En	Sv	LLaVA-NeXT	**0.00 (-0.01)**	**0.21 (-0.04)**	**0.04 (-0.02)**	**0.12 (-0.02)**	**0.81 (-0.01)**
		LLaVA-NeXT (FT)	0.00 (-0.07)	0.08 (-0.20)	0.02 (-0.11)	0.06 (-0.16)	0.78 (-0.07)
		Qwen-VL	0.00 (-0.01)	0.15 (-0.07)	0.02 (-0.03)	0.09 (-0.05)	0.79 (-0.03)
		Qwen-VL (FT)	0.00 (-0.06)	0.03 (-0.26)	0.01 (-0.11)	0.02 (-0.20)	0.76 (-0.08)
		mPLUG-Owl2	0.00 (-0.01)	0.14 (-0.11)	0.02 (-0.03)	0.09 (-0.07)	0.80 (-0.03)
		Phi-3	0.00 (-0.01)	0.05 (-0.14)	0.01 (-0.03)	0.04 (-0.08)	0.76 (-0.05)
		XComposer2	0.00 (-0.01)	0.11 (-0.13)	0.02 (-0.04)	0.08 (-0.07)	0.79 (-0.03)
En	Ru	LLaVA-NeXT	**0.00 (-0.01)**	**0.03 (-0.22)**	**0.00 (-0.05)**	**0.02 (-0.12)**	**0.89 (+0.07)**
		LLaVA-NeXT (FT)	0.00 (-0.07)	0.01 (-0.27)	0.00 (-0.13)	0.01 (-0.21)	0.72 (-0.13)
		Qwen-VL	0.00 (-0.01)	0.02 (-0.21)	0.00 (-0.05)	0.02 (-0.13)	0.85 (+0.03)
		Qwen-VL (FT)	0.00 (-0.06)	0.01 (-0.27)	0.00 (-0.12)	0.01 (-0.21)	0.70 (-0.14)
		mPLUG-Owl2	0.00 (-0.01)	0.01 (-0.23)	0.00 (-0.05)	0.01 (-0.14)	0.86 (+0.04)
		Phi-3	0.00 (-0.01)	0.01 (-0.19)	0.00 (-0.04)	0.01 (-0.12)	0.71 (-0.10)
		XComposer2	0.00 (-0.01)	0.02 (-0.22)	0.00 (-0.05)	0.02 (-0.13)	0.87 (+0.05)
En	Ja	LLaVA-NeXT	**0.01 (-0.00)**	**0.03 (-0.21)**	**0.01 (-0.05)**	**0.03 (-0.11)**	**0.84 (+0.03)**
		LLaVA-NeXT (FT)	0.00 (-0.07)	0.01 (-0.26)	0.00 (-0.13)	0.01 (-0.21)	0.73 (-0.12)
		Qwen-VL	0.00 (-0.01)	0.02 (-0.20)	0.00 (-0.05)	0.02 (-0.13)	0.83 (+0.00)
		Qwen-VL (FT)	0.00 (-0.06)	0.01 (-0.27)	0.00 (-0.12)	0.01 (-0.21)	0.72 (-0.12)
		mPLUG-Owl2	0.00 (-0.01)	0.02 (-0.23)	0.00 (-0.05)	0.02 (-0.14)	0.83 (+0.01)
		Phi-3	0.00 (-0.00)	0.02 (-0.18)	0.00 (-0.03)	0.02 (-0.11)	0.82 (+0.01)
		XComposer2	0.00 (-0.01)	0.02 (-0.22)	0.00 (-0.05)	0.02 (-0.12)	0.83 (+0.01)
En	Zh	LLaVA-NeXT	**0.00 (-0.01)**	**0.03 (-0.21)**	**0.01 (-0.05)**	**0.03 (-0.12)**	**0.83 (+0.01)**
		LLaVA-NeXT (FT)	0.00 (-0.07)	0.02 (-0.25)	0.01 (-0.12)	0.02 (-0.20)	0.73 (-0.12)
		Qwen-VL	0.00 (-0.01)	0.03 (-0.19)	0.01 (-0.04)	0.03 (-0.11)	0.83 (+0.01)
		Qwen-VL (FT)	0.00 (-0.06)	0.02 (-0.26)	0.00 (-0.11)	0.02 (-0.20)	0.72 (-0.12)
		mPLUG-Owl2	0.00 (-0.01)	0.02 (-0.22)	0.01 (-0.05)	0.02 (-0.14)	0.83 (+0.00)
		Phi-3	0.00 (-0.00)	0.02 (-0.18)	0.01 (-0.03)	0.02 (-0.11)	0.81 (-0.00)
		XComposer2	0.00 (-0.01)	0.03 (-0.21)	0.01 (-0.04)	0.03 (-0.12)	0.83 (+0.01)

Table 12: Other metrics results of LVLMs in Full Task. Bold fonts indicate the best score for that language combination. We also measured outputs with existing NLG (Sato et al., 2024) evaluation methods, BLEU (Kishore et al., 2002), ROUGE (Chin-Yew, 2004), and BertScore (Tianyi et al., 2019).

Input	Output	LVLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
			exact	partial		n=0	n=1	n=2	n=∞	
En	En	LLaVA-NeXT	26.49	31.54	26.07	1.35	1.65	1.66	1.70	252
		GPT-4o-mini	30.34	37.24	29.75	1.45	2.01	2.03	2.08	298
En	Es	LLaVA-NeXT	0.00	2.24	0.00	0.00	0.00	0.00	0.00	137
		GPT-4o-mini	14.32	22.78	15.43	1.20	1.50	1.80	1.95	153
En	Fr	LLaVA-NeXT	1.00	7.42	1.33	0.00	0.00	0.00	0.00	179
		GPT-4o-mini	18.89	21.56	19.74	0.80	1.40	1.60	1.72	192
En	De	LLaVA-NeXT	14.03	17.90	16.51	1.73	1.70	1.67	1.82	169
		GPT-4o-mini	18.65	26.78	19.89	1.95	2.01	2.13	2.25	181
En	It	LLaVA-NeXT	8.53	13.33	9.37	0.86	0.87	1.06	1.05	171
		GPT-4o-mini	15.45	18.76	14.12	0.95	1.20	1.45	1.68	176
En	Nl	LLaVA-NeXT	12.21	17.83	14.60	0.36	1.81	1.70	1.83	178
		GPT-4o-mini	15.89	21.34	16.78	1.30	1.60	1.78	1.90	184
En	Sv	LLaVA-NeXT	15.01	18.65	13.56	1.29	0.97	1.15	1.09	174
		GPT-4o-mini	16.32	23.47	17.65	1.35	1.50	1.65	1.80	181
En	Ru	LLaVA-NeXT	10.32	15.15	8.53	0.32	0.36	0.31	0.32	203
		GPT-4o-mini	12.89	17.65	13.45	0.65	0.90	1.10	1.25	210
En	Ja	LLaVA-NeXT	8.68	8.68	11.47	0.80	0.80	0.80	0.80	211
		GPT-4o-mini	16.23	18.32	18.56	0.95	1.10	1.20	1.32	225
En	Zh	LLaVA-NeXT	14.00	14.09	16.69	0.66	0.66	0.66	0.66	228
		GPT-4o-mini	17.45	21.09	18.89	1.23	1.40	1.56	1.70	239

Table 13: The result of Alignment-10 by gpt-4o-mini. This table focuses on En-{Lang}.

Input	Output	LVLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
			exact	partial		n=0	n=1	n=2	n=∞	
Es	Es	LLaVA-NeXT	17.26	21.30	17.05	2.13	2.32	2.17	2.10	186
		GPT-4o-mini	18.45	23.87	18.11	2.25	2.41	2.35	2.28	195
Fr	Fr	LLaVA-NeXT	24.35	29.27	24.38	0.95	0.90	0.88	0.90	211
		GPT-4o-mini	26.12	31.45	26.55	1.10	0.98	1.02	1.00	219
De	De	LLaVA-NeXT	17.45	20.66	21.05	2.11	2.20	2.22	2.11	204
		GPT-4o-mini	19.11	23.45	22.67	2.33	2.45	2.38	2.30	210
It	It	LLaVA-NeXT	10.34	15.43	11.33	1.16	0.93	0.96	0.96	185
		GPT-4o-mini	19.22	17.12	19.45	1.28	1.05	1.08	1.10	190
Nl	Nl	LLaVA-NeXT	17.66	23.56	19.78	0.79	3.55	3.61	3.88	199
		GPT-4o-mini	19.87	25.34	21.55	0.91	3.78	3.83	3.92	207
Sv	Sv	LLaVA-NeXT	27.51	29.61	16.71	2.10	0.87	0.89	0.90	206
		GPT-4o-mini	29.02	32.33	18.45	2.22	0.99	1.03	1.05	213
Ru	Ru	LLaVA-NeXT	14.38	17.43	9.81	0.26	0.45	0.42	0.41	219
		GPT-4o-mini	15.67	19.87	11.22	0.32	0.56	0.52	0.50	225
Ja	Ja	LLaVA-NeXT	13.38	13.38	17.68	0.73	0.83	0.83	0.83	249
		GPT-4o-mini	15.12	15.45	19.23	0.88	0.95	0.90	0.88	256
Zh	Zh	LLaVA-NeXT	13.78	13.78	17.00	0.54	0.53	0.53	0.53	246
		GPT-4o-mini	15.34	16.45	18.22	0.68	0.61	0.62	0.63	253

Table 14: The result of Alignment-10 by gpt-4o-mini. This table focuses on {Lang}-{Lang}.

I Comparison with OpenAI API

We also conducted Alignment-5 and Alignment-10 task by `gpt-4o-mini` (gpt-4o-mini-2024-07-18).

Input	Output	LVLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
			exact	partial		n=0	n=1	n=2	n=∞	
Es	En	LLaVA-NeXT	4.22	6.22	4.17	0.00	0.00	0.00	0.00	180
		GPT_4o-mini	25.23	27.85	25.02	1.80	1.80	1.82	2.10	224
Fr	En	LLaVA-NeXT	21.11	28.40	21.22	2.01	2.04	2.09	1.97	232
		GPT_4o-mini	22.43	30.12	23.08	2.22	2.25	2.32	2.18	238
De	En	LLaVA-NeXT	21.88	30.48	21.76	0.89	1.33	1.53	1.50	239
		GPT_4o-mini	23.01	32.15	23.56	1.12	1.45	1.60	1.55	244
It	En	LLaVA-NeXT	7.98	11.14	5.40	0.22	0.28	0.28	0.28	137
		GPT_4o-mini	12.44	14.75	10.23	1.35	1.40	1.42	1.40	142
Nl	En	LLaVA-NeXT	15.81	24.80	21.13	0.11	2.04	2.29	1.83	223
		GPT_4o-mini	17.02	26.45	22.75	0.23	2.30	2.55	2.01	230
Sv	En	LLaVA-NeXT	18.70	25.48	18.98	1.79	1.86	1.86	1.80	246
		GPT_4o-mini	23.85	27.89	20.45	1.95	2.02	2.05	1.95	252
Ru	En	LLaVA-NeXT	18.31	26.30	18.43	1.68	1.64	1.65	1.59	241
		GPT_4o-mini	24.50	28.12	19.22	1.85	1.80	1.83	1.75	247
Ja	En	LLaVA-NeXT	15.36	24.41	16.18	1.12	1.15	1.28	1.11	208
		GPT_4o-mini	18.85	26.12	17.89	1.28	1.35	1.40	1.32	215
Zh	En	LLaVA-NeXT	13.44	21.98	12.83	0.96	1.43	1.58	1.42	168
		GPT_4o-mini	18.98	23.56	15.12	1.11	1.61	1.75	1.60	172

Table 15: The result of Alignment-10 by gpt-4o-mini. This table focuses on {Lang}-En.

Input	Output	LVLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
			exact	partial		n=0	n=1	n=2	n=∞	
En	En	LLaVA-NeXT	22.58	30.30	25.19	3.15	2.30	2.40	2.44	259
En	En	GPT-4o-mini	29.60	35.40	28.10	3.40	2.50	2.60	2.55	259
En	Es	LLaVA-NeXT	5.76	8.21	7.39	0.38	0.34	0.35	0.32	160
En	Es	GPT-4o-mini	14.65	21.00	15.41	1.60	1.50	1.86	1.94	184
En	Fr	LLaVA-NeXT	6.85	12.94	7.03	0.22	0.16	0.17	0.17	181
En	Fr	GPT-4o-mini	19.71	21.01	20.12	1.24	1.18	1.19	1.18	182
En	Ja	LLaVA-NeXT	9.06	9.11	12.33	0.55	0.55	0.55	0.55	212
En	Ja	GPT-4o-mini	15.00	17.70	15.65	1.88	1.87	1.96	1.96	239
En	It	LLaVA-NeXT	4.10	7.54	5.19	0.11	0.07	0.06	0.09	118
En	It	GPT-4o-mini	17.15	18.85	18.61	1.13	1.08	1.07	2.10	193

Table 16: The result of Alignment-10 by gpt-4o-mini. This table focuses on En-{Lang}.

Input	Output	LVLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
			exact	partial		n=0	n=1	n=2	n=∞	
Es	Es	LLaVA-NeXT	12.77	18.26	10.97	1.29	1.22	1.22	1.08	203
Es	Es	GPT-4o-mini	16.89	21.42	18.11	1.31	1.26	1.25	1.12	223
Fr	Fr	LLaVA-NeXT	13.72	18.78	13.36	0.68	0.71	0.72	0.71	217
Fr	Fr	GPT-4o-mini	26.82	31.32	26.09	1.70	2.73	2.71	2.72	283
Ja	Ja	LLaVA-NeXT	20.68	20.68	24.14	1.04	1.06	1.06	1.06	247
Ja	Ja	GPT-4o-mini	23.73	23.72	27.10	1.02	1.07	1.07	1.08	214
It	It	LLaVA-NeXT	11.91	17.01	9.89	2.69	1.84	1.82	1.81	196
It	It	GPT-4o-mini	19.01	20.56	20.78	2.05	1.87	2.04	2.05	213

Table 17: The result of Alignment-10 by gpt-4o-mini. This table focuses on {Lang}-{Lang}.

Input	Output	LVLM	Entity Cov.		Entity F1	Entity Cooccurrence				Avg. Length
			exact	partial		n=0	n=1	n=2	n=∞	
Es	En	LLaVA-NeXT	29.15	33.34	29.88	0.35	0.32	2.05	2.01	220
Es	En	GPT-4o-mini	32.57	36.92	32.98	0.39	0.34	2.10	2.04	220
Fr	En	LLaVA-NeXT	3.57	5.33	3.27	0.00	0.00	0.00	0.00	162
Fr	En	GPT-4o-mini	22.80	29.10	26.60	1.32	1.51	1.51	1.61	193
Ja	En	LLaVA-NeXT	12.71	20.20	13.92	0.79	0.88	0.91	0.84	213
Ja	En	GPT-4o-mini	15.84	23.73	16.55	0.80	0.87	0.90	0.83	248
It	En	LLaVA-NeXT	9.95	17.66	10.50	0.66	0.42	0.39	0.38	219
It	En	GPT-4o-mini	13.10	17.50	14.02	1.68	1.89	1.99	1.99	298

Table 18: The result of Alignment-10 by gpt-4o-mini. This table focuses on {Lang}-En.

J Details of Each Language Templates

As indicated in Table 1, we created Templates for ten languages. Ten language templates are shown below. These templates were modified by nine native speakers of the country's first language, who were asked to modify the sentences to have the same nuance and level of detail as in English. We described these templates from Table 19 to Table 28.

Language	Type	Template
English	**Template 1**	
	Section	`Focus on {title} and explain the {section}.`
	Subsection	`In the context of {title}, explain the {subsection} and the {section}.`
	Sub subsection	`Focusing on the {section} of {title}, explain the {subsubsection} about the {subsection}.`
	Template 2	
	Section	`Explain the {section} of this artwork, {title}.`
	Subsection	`Explain the {subsection} about the {section} of this artwork, {title}.`
	Sub subsection	`Explain the {subsubsection} about the {subsection} of the {section} in this work, {title}.`
	Template 3	
	Section	`How does {title} explain its {section}?`
	Subsection	`In {title}, how is the {subsection} of the {section} explained?`
	Sub subsection	`Regarding {title}, how does the {section}'s {subsection} incorporate the {subsubsection}.`
	Template 4	
	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 19: Prompt Templates in English

Language	Type	Template
Japanese	**Template 1**	
	Section	{title}に焦点を当てて、その{section}を説明してください。
	Subsection	{title}の文脈で、{subsection}と{section}を説明してください。
	Sub subsection	{title}の{section}に焦点を当てて、{subsection}についての{subsubsection}を説明してください。
	Template 2	
	Section	{title}の{section}を説明してください。
	Subsection	{title}の{section}に関する{subsection}を説明してください。
	Sub subsection	{title}の{section}の{subsection}に関する{subsubsection}を説明してください。
	Template 3	
	Section	{title}はどのように{section}を説明していますか？
	Subsection	{title}では、どのように{section}の{subsection}が説明されていますか？
	Sub subsection	{title}に関して、{section}の{subsection}は{subsubsection}をどのように取り入れていますか？
	Template 4	
	Section	{title}に関して、どのように{section}が議論されていますか？
	Subsection	{title}の{section}における{subsection}の特徴を説明してください。
	Sub subsection	{title}の{section}について見たとき、その{subsection}の{subsubsection}をどのように議論しますか？

Table 20: Prompt Templates in Japanese

Language	Type	Template
Spanish	**Template 1**	
	Section	Concéntrate en {title} y explora la {section}.
	Subsection	En el contexto de {title}, explora la {subsection} y {section}.
	Sub subsection	Concentrándote en la {section} de {title}, explora la {subsubsection} sobre la {subsection}.
	Template 2	
	Section	Explora la {section} de esta obra de arte, {title}.
	Subsection	Explora la {subsection} sobre la {section} de esta obra de arte, {title}.
	Sub subsection	Explora la {subsubsection} sobre {subsection} de la {section} en esta obra de arte, {title}.
	Template 3	
	Section	¿Cómo aclara {title} su {section}?
	Subsection	En {title}, ¿Cómo se aclara la {subsection} de la {section}?
	Sub subsection	Con respecto a {title}, ¿Cómo la {subsection} de la {section} incorpora a la {subsubsection}?
	Template 4	
	Section	En {title}, ¿Cómo se discute la {section}?
	Subsection	Describe las características de la {subsection} en la {section} de {title}.
	Sub subsection	Al observar la {section} de {title}, ¿Cómo discutes la {subsubsection} de su {subsection}?

Table 21: Prompt Templates in Spanish

Language	Type	Template
Italian	**Template 1**	
	Section	Concentrati su {title} ed esplora la {section}.
	Subsection	Nel contesto di {title}, esplora la {subsection} e la {section}.
	Sub subsection	Concentrandosi sulla {section} di {title}, esplora la {subsubsection} sulla {subsection}.
	Template 2	
	Section	Esplora la {section} di questa opera d'arte, {title}.
	Subsection	Esplora la {subsection} sulla {section} di questa opera d'arte, {title}.
	Sub subsection	Esplora la {subsubsection} sulla {subsection} della {section} in questa opera, {title}.
	Template 3	
	Section	Come chiarisce {title} la sua {section}?
	Subsection	In {title}, come viene chiarita la {subsection} della {section}?
	Sub subsection	Per quanto riguarda {title}, come la {section} incorpora la {subsection} con la {subsubsection}?
	Template 4	
	Section	Come viene discussa la {section} in {title}?
	Subsection	Descrivi le caratteristiche della {subsection} nella {section} di {title}.
	Sub subsection	Osservando la {section} di {title}, come discuti la {subsection} della {subsubsection}?

Table 22: Prompt Templates in Italian

Language	Type	Template
French	**Template 1**	
	Section	Concentrez-vous sur {title} et expliquez la {section}.
	Subsection	Dans le contexte de {title}, expliquez la {subsection} et la {section}.
	Sub subsection	En vous concentrant sur la {section} de {title}, expliquez la {subsubsection} concernant la {subsection}.
	Template 2	
	Section	Expliquer la {section} de cette œuvre d'art, {title}.
	Subsection	Expliquer la {subsection} concernant la {section} de cette œuvre d'art, {title}.
	Sub subsection	Expliquer la {subsubsection} concernant la {subsection} de la {section} dans cette œuvre, {title}.
	Template 3	
	Section	Comment {title} explique-t-il sa {section}?
	Subsection	Dans {title}, comment la {subsection} de la {section} est-elle expliquée?
	Sub subsection	Concernant {title}, comment la {subsection} de la {section} intègre-t-elle la {subsubsection}?
	Template 4	
	Section	Dans {title}, comment est discutée la {section}?
	Subsection	Décrivez les caractéristiques de la {subsection} dans la {section} de {title}.
	Sub subsection	En examinant la {section} de {title}, comment discutez-vous la {subsubsection} de la {subsection}?

Table 23: Prompt Templates in French

Language	Type	Template
Chinese (Simplified)	**Template 1**	
	Section	专注于{title}并探索{section}。
	Subsection	在{title}的背景下，探索{subsection}和{section}。
	Sub subsection	专注于{title}的{section}，探索关于{subsection}的{subsubsection}。
	Template 2	
	Section	探索艺术作品{title}的{section}。
	Subsection	探索艺术作品{title}中关于{section}的{subsection}。
	Sub subsection	探索作品{title}中{section}的{subsection}的{subsubsection}。
	Template 3	
	Section	{title}是如何阐明其{section}的?
	Subsection	在{title}中，{section}的{subsection}是如何被阐明的?
	Sub subsection	关于{title}，{section}的{subsection}是如何结合{subsubsection}的?
	Template 4	
	Section	在{title}中，{section}是如何被讨论的?
	Subsection	描述{title}的{section}中{subsection}的特点。
	Sub subsection	在查看{title}的{section}时，你如何讨论其{subsection}的{subsubsection}?

Table 24: Prompt Templates in Chinese (Simplified)

Language	Type	Template
Dutch	**Template 1**	
	Section	Focus op {title} en leg de {section} uit.
	Subsection	In de context van {title}, leg de {subsection} en de {section} uit.
	Sub subsection	Gefocust op de {section} van {title}, leg de {subsubsection} over de {subsection} uit.
	Template 2	
	Section	Leg de {section} van dit kunstwerk uit, {title}.
	Subsection	Leg de {subsection} over de {section} van dit kunstwerk uit, {title}.
	Sub subsection	Leg de {subsubsection} over de {subsection} van de {section} in dit werk uit, {title}.
	Template 3	
	Section	Hoe verduidelijkt {title} zijn {section}?
	Subsection	Hoe wordt in {title} de {subsection} van de {section} verduidelijkt?
	Sub subsection	Met betrekking tot {title}, hoe incorporeert de {section}'s {subsection} de {subsubsection}?
	Template 4	
	Section	Hoe wordt de {section} besproken in {title}?
	Subsection	Beschrijf de kenmerken van de {subsection} in de {section} van {title}.
	Sub subsection	Wanneer je kijkt naar de {section} van {title}, hoe bespreek je de {subsection}'s {subsubsection}?

Table 25: Prompt Templates in Dutch

Language	Type	Template
Swedish	**Template 1**	
	Section	Fokusera på {title} och förklara {section}.
	Subsection	I samband med {title}, förklara {subsection} och {section}.
	Sub subsection	Med fokus på {section} i {title}, förklara {subsubsection} om {subsection}.
	Template 2	
	Section	Förklara {section} i detta konstverk, {title}.
	Subsection	Förklara {subsection} om {section} i detta konstverk, {title}.
	Sub subsection	Förklara {subsubsection} om {subsection} av {section} i detta verk, {title}.
	Template 3	
	Section	Hur förklarar {title} sitt {section}?
	Subsection	Hur förklaras {subsection} av {section} i {title}?
	Sub subsection	När det gäller {title}, hur innehåller {section}'s {subsection} {subsubsection}?
	Template 4	
	Section	I {title}, hur diskuteras {section}?
	Subsection	Beskriv egenskaperna hos {subsection} i {title}'s {section}.
	Sub subsection	När du tittar på {section} i {title}, hur diskuterar du dess {subsection}'s {subsubsection}?

Table 26: Prompt Templates in Swedish

Language	Type	Template
German	Section	Fokussiere dich auf {title} und erkunde erkläre die {section}.
	Subsection	Im Kontext von {title}, erkunde erkläre die {subsection} und die {section}.
	Sub subsection	Mit Fokus auf die {section} von {title}, erkunde erkläre die {subsubsection} über die {subsection}.
	Template 2	
	Section	Erkunde Erkläre die {section} dieses Kunstwerks, {title}.
	Subsection	Erkunde Erkläre die {subsection} über die {section} dieses Kunstwerks, {title}.
	Sub subsection	Erkunde Erkläre die {subsubsection} über die {subsection} der {section} in diesem Werk, {title}.
	Template 3	
	Section	Wie erläutert {title} seine {section}?
	Subsection	In {title}, wie wird die {subsection} der {section} erläutert?
	Sub subsection	Bezüglich {title}, wie integriert die {subsection} der {section} die {subsubsection}?
	Template 4	
	Section	Wie wird die {section} in {title} diskutiert?
	Subsection	Beschreibe die Merkmale der {subsection} in der {title}'s {section}.
	Sub subsection	Wenn du die {section} von {title} betrachtest, wie diskutierst du die {subsection}'s {subsubsection} von der {subsection}?

Table 27: Prompt Templates in German

Language	Type	Template
Russian	Section	Сосредоточьтесь на {title} и объясните {section}.
	Subsection	В контексте {title} объясните {subsection} и {section}.
	Sub subsection	Сосредоточившись на {section} в {title}, объясните {subsubsection} о {subsection}.
	Template 2	
	Section	Объясните {section} этого произведения искусства, {title}.
	Subsection	Объясните {subsection} о {section} этого произведения искусства, {title}.
	Sub subsection	Объясните {subsubsection} о {subsection} в {section} этого произведения, {title}.
	Template 3	
	Section	Как {title} объясняет свой/свою {section}?
	Subsection	Как объясняется в {title} {subsection} в {section}?
	Sub subsection	Что касается {title}, как {section} в {subsection} включает {subsubsection}?
	Template 4	
	Section	Как обсуждается {section} в {title}?
	Subsection	Опишите черты {subsection} в {section} в {title}.
	Sub subsection	Когда вы рассматриваете {section} в {title}, как вы обсуждаете {subsubsection} в {subsection}?

Table 28: Prompt Templates in Russian

K Dataset Example

The below block is an example of datasets we created. The full dataset is available at https://huggingface.co/datasets/naist-nlp/MultiExpArt.

Dataset Example


```
{
"text_id": 1,
"title": "Mona Lisa",
"en_title": "Mona Lisa",
"lang": "en",
"source_lang": "en",
"target_lang": "en",
"image_url": "IMAGE_URL_FOR_MONA_LISA",
"en_image_url": "IMAGE_URL_FOR_MONA_LISA",
"template": "en_template1_section",
"prompt": "Focus on Mona Lisa and explain the Abstract.",
"gold": "The Mona Lisa is a half-length portrait painting by Italian artist Leonardo da Vinci. Considered an archetypal masterpiece of the Italian Renaissance, it has been described as the best known, the most visited, the most written about, the most sung about, [and] the most parodied work of art in the world. The painting's novel qualities include the subject's enigmatic expression, monumentality of the composition, the subtle modelling of forms, and the atmospheric illusionism. The painting has been traditionally considered to depict the Italian noblewoman Lisa del Giocondo. It is painted in oil on a white poplar panel. Leonardo never gave the painting to the Giocondo family.",
"gold_entity": ["Francis I of France", "Italian Renaissance", "Italy", "Francis I of France", "Leonardo da Vinci", "Lisa del Giocondo", "Louvre", "Louvre", "Nat King Cole"]
}
```

L Examples of LVLM outputs

Tables 29 and 30 present output examples from LLaVA-NeXT.

Question	Input	Output	LVLM output
Focus on Venus de Milo and explain the Description.	En	En	The image you've provided appears to be of the Venus de Milo, a famous marble sculpture of the Greek goddess Aphrodite (Venus in Roman mythology). The sculpture is known for its missing arms, which has led to various interpretations and speculations about its original form and the story it was meant to tell. The Venus de Milo is characterized by its draped clothing, which is typical of classical Greek sculpture, and the flowing drapery that suggests movement. The figure is depicted in a standing pose, which is common for depictions of the goddess, and the facial expression is serene and contemplative.The sculpture is missing its arms, which has led to much speculation about the original intent of the artist. Some believe that the missing arms were removed to prevent the statue from being used for religious purposes, while others think that the arms were removed for artistic reasons, to emphasize the beauty of the torso and the elegance of the pose.The Venus de Milo is one of the most celebrated sculptures in the world and is housed in the Louvre Museum in Paris, France. It is a symbol of beauty and has been the subject of numerous reproductions and adaptations throughout history.
	En	Ja	Venus de Miloは、ギリシャの古代彫刻で有名な女性像の一つで、世界中で愛される名彫刻の一つです。Venus de Miloは、ギリシャの島であるミロス島で発見されましたが、その後、様々な謎に包まれています。2015年4月25日から9月6日まで、三菱地所美術館で「Venus de Miloとその謎」という展覧会が開催されます。この展覧会では、Venus de Miloの謎を解くための最新の科学技術が活用され、Venus de Miloの美しさと魅力を再び発見することができます。Venus de Miloは、ギリシャの島であるミロス島で発見されました。
	En	Ja (DeepL)	Venus de Milo is one of the most famous female figures in ancient Greek sculpture, and one of the most beloved masterpieces of sculpture in the world.Venus de Milo was discovered on the Greek island of Milos, but has since been shrouded in mystery.From April 25 to September 6, 2015, the Mitsubishi Jisho Art Museum will host the exhibition "Venus de Milo and its Mysteries". The exhibition will utilize the latest scientific techniques to solve the mystery of Venus de Milo, allowing visitors to rediscover the beauty and charm of Venus de Milo, which was discovered on the Greek island of Milos.

Table 29: Comparison of output examples from LLaVA-NeXT. (DeepL) represents translations from other languages into English.

Question	Input	Output	LVLM output
Focus on Venus de Milo and explain the Description.	Ja	Ja	この彫刻は、ローマのサン・ピエトロ大聖堂の前庭に立つ。ミロのヴィーナスは、ローマのサン・ピエトロ大聖堂の前庭に立つ。
	Ja	Ja (DeepL)	This sculpture stands in the vestibule of St. Peter's Basilica in Rome. The Venus de Milo stands in the vestibule of St. Peter's Basilica in Rome.
	Ja	En	The statue is a nude, with a contrapposto pose. The statue is 1.5 meters tall, and is 0.8 meters wide. The statue is on a marble base that is 0.8 meters tall, and 0.9 meters wide. The statue is on display in the Louvre Museum.
	–	Gold	The Venus de Milo is an over tall. Brill's New Pauly says . The Louvre's online catalogue states . Christofilis Maggidis says . Alain Pasquier says excluding the plinth, or including it. Parian marble statue of a Greek goddess, most likely Aphrodite, depicted with a bare torso and drapery over the lower half of her body. The figure stands with her weight on her right leg, and the left leg raised; her head is turned to the left. The statue is missing both arms, the left foot, and the earlobes. There is a filled hole below her right breast that originally contained a metal tenon that would have supported the right arm. The Venus' flesh is polished smooth, but chisel marks are still visible on other surfaces. The drapery is more elaborately carved on the right-hand side of the statue than the left, perhaps because on the left-hand side it was originally obscured from view. Likewise the Venus is less finely-finished from behind, suggesting that it was originally intended to be viewed only from the front. While the body of the Venus is depicted in a realistic style, the head is more idealised. The lips are slightly open, the eyes and mouth are small. The sculpture has been minimally restored: only the tip of the nose, lower lip, big toe on the right foot, and some of the drapery.Stylistically, the sculpture combines elements of classical and Hellenistic art. Features such as the small, regular eyes and mouth, and the strong brow and nose, are classical in style, while the shape of the torso and the deeply carved drapery are Hellenistic.Kenneth Clark describes the figure as "the last great work of antique Greece", and "of all the works of antiquity one of the most complex and the most artful. ...[the sculptor] has consciously attempted to give the effect of a 5th-century work", while also using "the inventions of his own time"; "the planes of her body are so large and calm that at first we do not realise the number of angles through which they pass. In architectural terms, she is a baroque composition with classic effect".

Table 30: Comparison of output examples from LLaVA-NeXT. (DeepL) represents translations from other languages into English.

M Pre-trained data list

This study specifies the pretraining data used for the open models employed. As of September 2024, Phi-3 and LLaVA-NeXT have not been released.

Data Type	**Data Name**	**Model**		
		mPLUG-Owl2 (Qinghao et al., 2024)	**Qwen-VL** (Jinze et al., 2023b)	**XComposer2** (Xiaoyi et al., 2024)
Text	ShareGPT (Lin et al., 2025)	✓	✓	✓
	SlimOrca (Lian et al., 2023)	✓		
	In-house Data	✓	✓	✓
Dialogue Caption	LLAVA (Haotian et al., 2023b)	✓		
	COCO (Xinlei et al., 2015)	✓	✓	✓
	TextCaps (Oleksii et al., 2020)	✓		✓
	SBU (Vicente et al., 2011)		✓	
	DataComp (Schmidt, 2023)	✓		
	CC12M & 3M (Soravit et al., 2021)	✓		✓
	LAION-en & zh (Schuhmann et al., 2022)	✓		✓
VQA	VQA2 (Goyal et al., 2017)	✓	✓	
	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 (Antol et al., 2015)	✓		
Grounding	A12D	✓		
	Ref Grounding		✓	
	GRIT (Peng et al., 2023)	✓		
	VisualGenome (Krishna et al., 2016)	✓	✓	
	RefCOCO (Zarrieß and Schlangen, 2018)		✓	
	RefCOCO+ (Zarrieß and Schlangen, 2018)		✓	
	RefCOCOg		✓	
OCR	SynthDoG-en & zh (Kim et al., 2022)	✓		
	Common Crawl pdf & HTML	✓		
Image Captioning	Web CapFilt (Li et al., 2022)	✓	✓	
	NoCaps	✓		✓
Visual Spatial Reasoning	Flickr30K (Hambardzumyan et al., 2022)		✓	✓
	IconQA (Lu et al., 2021)			
Video Question Answering	MSVD-QA		✓	
	MSRVT-QA	✓		
	iVQA (Liu et al., 2018)		✓	
Image Classification	VizWiz (Gurari et al., 2018)		✓	
Knowledge-Grounded Image QA	ScienceQA (Lu et al., 2022)		✓	

Table 31: Pretraining dataset list.