Appendix Material

1 Re-evaluation on VSR

We first re-evaluated the VLLM from scratch and diagnosed issues of inconsistent performance, hypersensitiveness on text prompts, lack of perception on visual details and answer bias. The absence of a unified instruction test set has resulted in significant variance in model performance. The instability of instruction-following hinders the evaluation and optimization of VLLM's VSR capabilities. Therefore, we proposed a unified instruction test set by expanding the VSR test set through both manual and GPT4-generated templates.

1.1 Variance and In-consistency on VSR

We first provide the most comprehensive summary including over 120 attempts(including Prism's results (Karamcheti et al., 2024)) across more than 30 models in Table 1 and Table 2.

First, we tested GPT-4o using the third template in Table 3 and achieved the highest results 84.6%. However, upon further analysis, we found that the model's performance exhibited significant variance. Most VLLMs had an accuracy slightly better than the random guess rate of 50%, while some could reach up to 70%, particularly the "Test only LMs." These results raise serious doubts about the effectiveness of current VLLMs in VSR tasks.

During the summary constitution, we also identified inconsistencies in model performance, such as the model LLaVA1.5 having nearly a 20% lower accuracy in MiniGPTv2's reproduction(51%)(Chen et al., 2023) compared to the result in Prism(71%)(Karamcheti et al., 2024).

We suspect that this phenomenon is due to the lack of a unified question instruction test set. Different studies may have selectively used question templates that favored their own models when making comparisons. Combined with the possibility that models are extremely sensitive to text input, this could have led to the drastic inconsistencies in results. To validate our hypothesis, we first designed manual templates for testing.

1.2 Manual Templates

To validate our hypothesis, we meticulously designed our prompt templates through four distinct phases:

1. In the first phase, we used the simplest questioning style, avoiding any unnecessary words and covering most binary question formats in Table 3. Here, the [caption] represents the combination of the triplet [subject] [relation] [object].

2. In the second phase, in Table 4, we add phrases like "in the image", "base on the image" or "focus on the image" to prompt the model to concentrate on visual information.

3. In the third phase, in Table 5, we made our prompts more concrete and specific, add prompts using words as "Distinguish the positional relation", "Focus on the positional relation" or "Reason on the spatial relation" to encourage model to be absorbed in spatial reasoning.

4. In the last fourth phase, in Table 6, we further provided the ground truth [subject box] and [object box] to assist the model in its reasoning process.

We tested all the templates on LLaVA1.5 13B(the most common used VLLM structure), and the results, along with the templates, are displayed in Tables 3, 4, 5, and 6. Since generative models may not always fully adhere to the instructions when answering questions, We expanded the original accuracy metric to include the proportion of instances where the model successfully followed the instructions when answering. On this basis, we further tested the accuracy of the answers. In the result

VLMM	Version	LLM	Vision Encoder	ACC(%)
GPT-4o	our template			84.6
Bard (Shao et al., 2023a)	*	*	*	82.0
GPT-4V (Lei et al., 2024)	naive	-	-	67.8
	CoT	-	-	70.4
	Scaffolding Coordinates	-	-	74.4
mPLUG-OWL (Ye et al., 2024)		LLaMA-7B	CLIP	46.0
OpenFlamingo-V2 (Awadalla et al., 2023)	*	*	*	58.0
Otter	*	*	*	24.0
Otter-I (Li et al., 2023a)	*	*	*	56.0
VPGTrans (Zhang et al., 2023a)	*	*	*	40.0
PandaGPT (Su et al., 2023)		vicuna-13B	Imagebind	46.7
LLaMA-Adapter (Zhang et al., 2023b)	LA	LLaMA 7B	CLIP (Multi-scale)	50.6
	LA v2	LLaMA 7B	CLIP (Multi-scale) + caption expert	52.0
Cobra (Zhao et al., 2024)		Mamba-2.8B	DINOv2+SigLIP	63.6
Mini-GPT4 (Zhu et al., 2023)	13B	LLaMA2-chat (7B)	EVA	41.6
	7B			60.6
	7B chat			62.9
LLaVA (Liu et al., 2023)	v1.5	vicuna-7B	CLIP	51.4
		vicuna-13B	CLIP	51.2
InstructBLIP (Dai et al., 2023)		FlanT5XL	CLIP + QF	64.8
		FlanT5XXL		65.6
		vicuna-7B		54.3
		vicuna-13B		52.1
BLIP2 (Li et al., 2023b)		FlanT5XL	CLIP + QF	60.5
		FlanT5XXL		68.2
		vicuna-7B		50.0
		vicuna-13B		50.9
ImageBind-LLM (Han et al., 2023)		LLaMA	imagebind	49.3
	(D)			49.7
Qwen-VL (Bai et al., 2023)	Qwen7B	Qwen7B	ViT-bigG	63.8
	Qwen7B-chat			67.5
SPHINX (Lin et al., 2023)		LLaMA 2	mixed*	62.6
	1k			62.9
	2k			63.1

Table 1: Part 1 Summary of popular VLMMs' performance on VSR dataset. In the first row, we evaluated the GPT-4o with it's API using our own prompt mentioned later in our *Test-S* set. Label "*" are result gathered from LVLM-eHub(Shao et al., 2023b) and "-" are from *SCAFFOLD*(Lei et al., 2024) release.

Models	Version	LLM	Vision Encoder	ACC(%)
Prismer (Liu et al., 2024)		RoBERTa	CLIP	68.4
	+ Normal			68.3
	+ Edge			67.8
	+ Seg.			68.4
	+ OCR Det.			67.2
	+ Obj. Det.			68.3
	No Experts			65.6
	+ 6 Experts			68.7
Prism* (Karamcheti et al., 2024)	Prism-CLIP 7B (C)	LLaMA 2 7B	CLIP	66.6
	Prism-CLIP 7B			57.7
	Prism-SigLIP 7B (C)		SigLIP	65.1
	Prism-SigLIP 7B			56.7
	Prism-DINOSigLIP 7B (C)		DINO+SigLIP	66.2
	Prism-DINOSigLIP 7B			59.5
	Prism-CLIP 13B (C)	LLaMA 2 13 B	CLIP	65.9
	Prism-CLIP 13B			71.8
	Prism-SigLIP 13B (C)		SigLIP	62.8
	Prism-SigLIP 13B			64.5
	Prism-DINOSigLIP 13B (C)		DINO+SigLIP	71.8
	Prism-DINOSigLIP 13B			72.1
BLIVA (Hu et al., 2023)		vicuna13B	CLIP + QF	62.2
		FlanT5XXL		68.8
VisLingInstruct (Zhu et al., 2024)		FlanT5XL	EVA	64.1
		FlanT5XXL		66.9
		vicuna-7B		60.1
		vicuna-13B		56.2
Text-only LMs (Azkune et al., 2024)		BERT-base (110M)		73.6
		BERT-large (336M)		74.4
		T5-base (220M)		73.1
		T5-large (770M)		74.4
		T5-3B (3B)		74.5
VisualBERT (Li et al., 2019)				51.0
LXMERT (Tan and Bansal, 2019)				61.2
VILT (Kim et al., 2021)				63.0

Table 2: Part 2 Summary of popular Models performance on VSR dataset including VLLMs, Text-only LMs and traditional pre trained VLMs. Prism (Karamcheti et al., 2024) investigated the design space of visually-conditioned language models and provide more 55 results in its last page of the appendix.

Tables, "answered" represents the proportion of successfully following the question instructions, regardless of whether the answers are correct. It reflects the model's ability to adhere to instructions in binary questions, "acc_t" represents the overall answer accuracy totally, and "acc_a" refers to the accuracy rate of the questions that were answered correctly based on the amount of answered questions. Here is how they are calculated:

$$answered = \frac{Amount_{answered}}{Amount_{total}}$$

$$aac_t = \frac{Amount_{correct}}{Amount_{total}}$$

$$aac_a = \frac{Amount_{correct}}{Amount_{answered}}$$

To provide a more intuitive comparison, we illustrated the changes in various accuracy metrics as the templates were progressively enriched through each phase in Figure 1.

1.3 Hyper-senstivity on Language Prompt

Although our manually designed templates may not be exhaustive and might not cover all real-world visual spatial reasoning scenarios, our limited set of templates is sufficient to reveal the issue of models being overly sensitive to text prompts.

As shown in Line Chart in Figure1, the accuracy metrics exhibit significant fluctuations as the template number increases. Adjacent templates can result in accuracy differences of up to 15%, even though the only change between these templates might be just one or two words.

Specifically, the way that questions are phrased, along with the addition, removal, or substitution of meaningless words and phrases, as well as the order in which these words are positioned, all significantly impact the model's final accuracy in answering the questions. For example, comparing template 2 and template 3 Table 3, simply replacing the phrase "True or not." at the end of the sentence with "True or false." led to an over 20% increase in the overall accuracy of the model's responses.

Additionally, we found that the model does not fully understand the semantics of words like "Spatial" and "Positional". In phrase 3 Table 5, we specifically included such words to guide the model towards focusing more on spatial information and to enhance its reasoning based on this focus. However, the results were disappointing. After adding these guiding words, the model became even more confused, and the accuracy of its responses dropped sharply. Most values fell below 50%, which is worse than random guess. This confirms that the model does not understand the specific meaning of terms like "Spatial" and is even less likely to connect them with specific visual features in an image.

Finally, in phase 4 Table6, we incorporated COCO-style bounding box information. Although LLaVA's fine-tuning data also included bounding boxes, and the model had been exposed to such data before, the accuracy did not improve as expected.

Overall, when processing text inputs, the accuracy of the model's responses is significantly affected by the way questions are phrased. Even minor changes in the wording of the input can notably impact the model's reasoning ability. Additionally, it was observed that the more concise and straightforward the text input, the less extraneous information the model needs to handle, leading to relatively higher response accuracy.

1.4 Under-Senstivity on Vision Details

Given that the model exhibits hyper-sensitivity to language prompts, it raises concerns about its ability to accurately discern positional information in the visual domain. This suggests that the model's reliance on textual input may overshadow its capability to interpret and reason about visual spatial relationships effectively.

Indeed, a robust visual spatial reasoning model must accurately identify different visual positional information. Within the current mainstream VLLM architecture, a key requirement is that the visual backbone extracts and aligns visual tokens with significant discriminative power in the "positional" dimension. This ensures that the model can effectively differentiate between various spatial relationships in the visual data, which is crucial for reliable reasoning and accurate responses in such VSR tasks.

In Figure 2, we arrange a case study and plot a scatter chart of 200 samples across each 7 common spatial relations to illustrate under-sensitivity to vision details. The 7 common spatial relations include "above", "next to", "behind", "inside", "on top of", "under" and "on". To begin, we perform average pooling on the visual tokens along the token length dimension. This operation reduces the dimensionality of the tokens, making them more manageable for further analysis. Next, we use t-

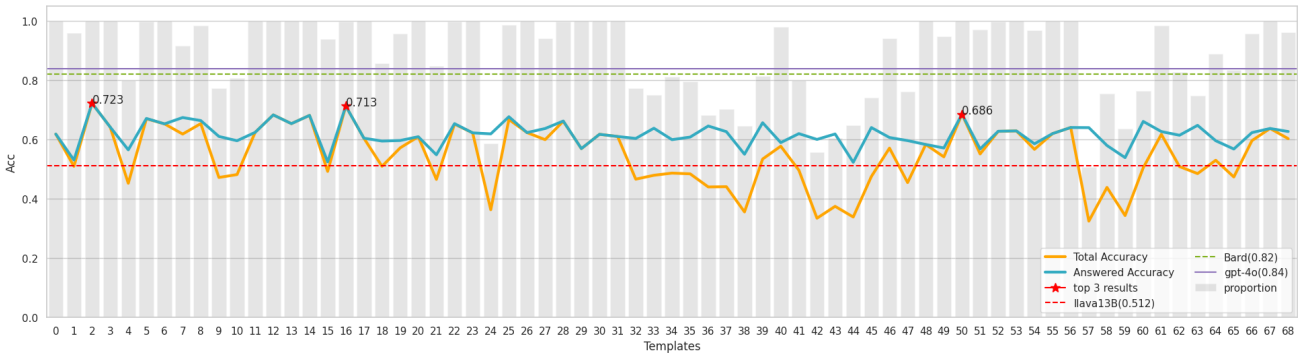

Figure 1: Result of LLaVA1.5 13B on VSR across 69 manually generated templates. The gray bar is the answered proportion on each template. The orange line indicate the total accuracy and green represented the answered accurac. The top 3 templates are labeled by red star. For comparison, we have also plotted some of the summarized results. The best-performing GPT-4o model is marked with a purple line and label, the second-best BARD is indicated with a green dashed line, and the results for the tested LLaVA 1.5 13B model are shown with a red dashed line.

Index	Templates of phase 1(the simplest questioning style)	answered	acc_t	acc_a
1	[caption], Yes or no.	1.00	0.61	0.61
2	[caption], True or not.	0.96	0.50	0.53
3	[caption], True or false.	1.00	0.72	0.72
4	Whether the [subject] is [relation] [object]? Yes or no.	1.00	0.64	0.64
5	Whether the [subject] is [relation] [object]? True or not.	0.80	0.45	0.56
6	Whether the [subject] is [relation] [object]? True or false.	1.00	0.67	0.67
7	Is the [subject] [relation] [object]? yes or no.	1.00	0.65	0.65
8	Is the [subject] [relation] [object]? True or not.	0.91	0.61	0.67
9	Is the [subject] [relation] [object]? True or false.	0.98	0.65	0.66
10	Whether [caption]? Answer the question with yes or no.	0.77	0.47	0.61
11	Is the [subject] [relation] [object]? Answer the question with yes or no.	0.80	0.48	0.59
12	Whether [caption]? A. yes B. no Answer with the option's letter from the given choices directly.	1.00	0.62	0.62
13	Is the [subject] [relation] [object]? A. yes B. no Answer with the option's letter from the given choices directly.	1.00	0.68	0.68
14	Answer the following binary question with the capital letter of the answer list below. Whether [caption]? A.yes B.no.	1.00	0.65	0.65
15	Answer the following binary question with the capital letter of the answer list below. Is the [subject] [relation] [object]? A.yes B.no.	1.00	0.68	0.68

Table 3: Templates of phase 1 which use the simplest questioning style avoiding any unnecessary words and covering most binary question formats. This includes two types of questioning methods: general declarative statements and general interrogative sentences. It covers questions formulated with "is" and "whether", as well as multiple-choice question formats. The replaceable parts in the templates are highlighted in colored fonts.

Index	Templates of phase 2 (focus on image side)	answered	acc_t	acc_a
16	[caption] in the image, True or not.	0.93	0.49	0.52
17	[caption] in the image , True or not.	1.00	0.71	0.71
18	[caption] in the image , True or false.	1.00	0.60	0.60
19	In the image, [caption], True or not.	0.85	0.50	0.59
20	In the image, [caption], True or false.	0.95	0.57	0.59
21	In the image, [caption], Yes or no.	1.00	0.60	0.60
22	Base on the image, [caption], True or not.	0.84	0.46	0.54
23	Base on the image, [caption], True or false.	1.00	0.65	0.65
24	Base on the image, [caption], Yes or no.	1.00	0.62	0.62
25	This is an image about [caption], True or not.	0.58	0.36	0.61
26	This is an image about [caption], True or false.	0.98	0.66	0.67
27	This is an image about [caption], Yes or no.	1.00	0.62	0.62
28	There is a [subject] [relation] the [object] in the image, True or not.	0.94	0.59	0.63
29	There is a [subject] [relation] the [object] in the image, True or false.	1.00	0.66	0.66
30	There is a [subject] [relation] the [object] in the image, Yes or no.	1.00	0.56	0.56
31	Whether [caption] in the image? yes or no.	1.00	0.61	0.61
32	Is the [subject] [relation] the [object] in the image? yes or no.	1.00	0.61	0.61
33	Whether [caption]? Focus on the image and answer yes or no.	0.77	0.46	0.60
34	Is the [subject] [relation] the [object]? Focus on the image and answer yes or no.	0.75	0.47	0.63
35	Focus on the image and answer the question with yes or no, Whether [caption]?	0.81	0.48	0.59
36	Focus on the image and answer the question with yes or no, Is the [subj] [relation] the [object]?	0.79	0.48	0.60

Table 4: Templates of phase 2 which add phrases like “in the image”, “base on the image” or “focus on the image” to prompt the model to concentrate on visual information. We placed these phrases at both the beginning and the end of our original questions. The replaceable parts in the templates are highlighted in colored fonts.

index	Templates of phase 3 (focus on positional&spatial relations)	answered	acc_t	acc_a
37	Distinguish the positional relation in the image, whether [caption]?	0.68	0.44	0.64
38	Focus on the positional relation in the image, whether [caption]?	0.70	0.44	0.62
39	Reason on the spatial relation in the image, whether [caption]?	0.64	0.35	0.55
40	Distinguish the positional relation in the image, is the [subject] [relation] the [object]?	0.81	0.53	0.65
41	Focus on the positional relation in the image, is the [subject] [relation] the [object]?	0.97	0.57	0.58
42	Reason on the spatial relation in the image, is the [subject] [relation] the [object]?	0.80	0.49	0.62
43	Distinguish the positional relation between the [subject] and the [object] in the image, whether [caption]?	0.55	0.33	0.60
44	Focus on the positional relation between the [subject] and the [object] in the image, whether [caption]?	0.60	0.37	0.61
45	Reason on the spatial relation between the [subject] and the [object] in the image, whether [caption]?	0.64	0.33	0.52
46	Distinguish the positional relation between the [subject] and the [object] in the image, is the [subject] [relation] the [object]?	0.74	0.47	0.64
47	Focus on the positional relation between the [subject] and the [object] in the image, is the [subject] [relation] the [object]?	0.94	0.57	0.60
48	Reason on the spatial relation between the [subject] and the [object] in the image, is the [subject] [relation] the [object]?	0.76	0.45	0.59

Table 5: Templates of phase 3 that focus more on positional&spatial relations. We made our prompts more concrete and specific by adding prompts using words such as "Distinguish the positional relation", "Focus on the positional relation" or "Reason on the spatial relation" to encourage model to be absorbed in spatial reasoning.

Index	Templates of phase 4 (bounding box added)	answered	acc_t	acc_a
49	The [subject] [subject box] is [relation] [object] [object box], yes or no.	1.00	0.58	0.58
50	The [subject] [subject box] is [relation] [object] [object box], True or not.	0.94	0.54	0.57
51	The [subject] [subject box] is [relation] [object] [object box], True or false.	1.00	0.68	0.68
52	Whether the [subject] [subject box] is [relation] [object] [object box]? Yes or no.	0.97	0.55	0.56
53	Whether the [subject] [subject box] is [relation] [object] [object box]? True or not.	1.00	0.62	0.62
54	Whether the [subject] [subject box] is [relation] [object] [object box]? True or false.	1.00	0.62	0.62
55	Is the [subject] [subject box] [relation] [object] [object box]? yes or no.	0.96	0.56	0.58
56	Is the [subject] [subject box] [relation] [object] [object box]? True or not.	1.00	0.62	0.62
57	Is the [subject] [subject box] [relation] [object] [object box]? True or false.	1.00	0.64	0.64
58	Distinguish the positional relation between the [subject] [subject box] and the [object] [object box] in the image, whether [caption]?	0.50	0.32	0.64
59	Focus on the positional relation between the [subject] [subject box] and the [object] [object box] in the image, whether [caption]?	0.75	0.43	0.58
60	Reason on the spatial relation between the [subject] [subject box] and the [object] [object box] in the image, whether [caption]?	0.63	0.34	0.53
61	Distinguish the positional relation between the [subject] [subject box] and the [object] [object box] in the image, is the [subject] [relation] the [object]?	0.76	0.50	0.66
62	Focus on the positional relation between the [subject] [subject box] and the [object] [object box] in the image, is the [subject] [relation] the [object]?	0.98	0.61	0.62
63	Reason on the spatial relation between the [subject] [subject box] and the [object] [object box] in the image, is the [subject] [relation] the [object]?	0.82	0.50	0.61
64	Distinguish the positional relation between the [subject] and the [object] in the image, Bounding box list: [subject] [subject box] , [object] [object box], whether [caption]?	0.74	0.48	0.64
65	Focus on the positional relation between the [subject] and the [object] in the image, Bounding box list: [subject] [subject box] , [object] [object box], whether [caption]?	0.88	0.53	0.59
66	Reason on the spatial relation between the [subject] and the [object] in the image, Bounding box list: [subject] [subject box] , [object] [object box], whether [caption]?	0.83	0.47	0.56
67	Distinguish the positional relation between the [subject] and the [object] in the image, Bounding box list: [subject] [subject box] , [object] [object box], is the [subject] [relation] the [object]?	0.95	0.59	0.62
68	Focus on the positional relation between the [subject] and the [object] in the image, Bounding box list: [subject] [subject box] , [object] [object box], is the [subject] [relation] the [object]?	1.00	0.63	0.63
69	Reason on the spatial relation between the [subject] and the [object] in the image, Bounding box list: [subject] [subject box] , [object] [object box], is the [subject] [relation] the [object]?	0.96	0.60	0.62

Table 6: Templates of phase 4 with bounding box of related entities added. We further provided the ground truth [subject box] and [object box] to assist the model in its reasoning process.

SNE (t-Distributed Stochastic Neighbor Embedding) from sklearn to further reduce the dimensionality of these pooled features. T-SNE is particularly useful for visualizing high-dimensional data in a two- or three-dimensional space. Finally, we plot the results as a scatter plot, which helps us visualize how the visual tokens are distributed and whether distinct positional relationships are well-separated in this lower-dimensional space. This process allows us to assess the effectiveness of the visual backbone in capturing and differentiating positional information.

The results indicate a serious overlapping phenomenon that the visual tokens corresponding to different positional categories are mixed together, making it difficult to distinctly separate them. This suggests that the visual tokens fed into the LLM (Large Language Model) are inherently under-sensitive to positional information.

In other words, the visual backbone may not be effectively encoding or preserving the critical spatial relationships needed for accurate visual-spatial reasoning. This under-sensitivity could be a significant factor to the model's challenges in accurately distinguishing positional relationships, leading to suboptimal performance in visual-spatial tasks.

1.5 Bias on Binary QA Senario

The model exhibited severe response biases for most template answers. In Histogram Figure3, we illustrate the comparison of model accuracy on "yes" (green) and "no" (red) questions across 41 templates, where the response answered rate exceeds 0.85. The results reveal a significant imbalance, with the accuracy of "yes" answers consistently higher than "no" answers across most templates. This raises concerns that the relatively high accuracy might be due to the model's tendency or shortcut to default to generating "yes" answers, rather than truly understanding and reasoning about the content.

We suspect this may be due to the co-occurrence of subject and object concepts in both question text and images. When posing questions, the input text includes entities that are present in the image. This co-occurrence phenomenon may confuse the model, leading it to hastily provide "yes" answers based on co-occurrence rather than focusing on the actual visual spatial relationship. The model must focus on the spatial relationships between these entities, rather than relying on the mere co-occurrence of object concepts in both the text and the image. This distinction is crucial for tasks requiring precise visual-spatial reasoning.

Compared to "yes" questions, correctly identifying the spatial relationships in "no" questions is more challenging and provides a better reflection of the model's visual reasoning capabilities. This is because "no" answers often require the model to discern subtle or complex differences in the spatial arrangement of entities, making it a more rigorous test of the model's ability to understand and reason about visual information accurately. This finding highlights a potential limitation in the model's ability to handle binary questions objectively and may indicate a need for further refinement to address this bias and improve the model's overall reliability in diverse question scenarios.

2 Image Controllable Expansion

We discuss the image controllable expansion in this section. Diffusion models for controlled image generation offer a powerful and flexible approach to creating images that meet specific criteria. By incorporating conditioning information into the denoising process, models can generate highly controlled and precise outputs. They represent a powerful approach for image repainting, leveraging the systematic denoising process to generate high-quality, contextually consistent images. In this work, we utilized the most 3 popular applications to expand the image data by SDXL as follows:

2.1 Text to Image

It generates an image from a text description. The denoising process is guided by the text, and once the denoising process ends after a predetermined time steps, the image representation is decoded into an image.

In Figure 4, we showcase some of the repaint results. It can be observed that, while the spatial relationships between the main entities remain unchanged, the generated images exhibit greater diversity. This includes not only changes in details but also variations in the appearance of the main entities, the overall color scheme, and even the artistic style of the images.

However, it is worth noting that the newly generated images may occasionally exhibit detail issues and entity hallucination problems due to the lack of guidance from the original images. For example, the arm detail of the person in the third column

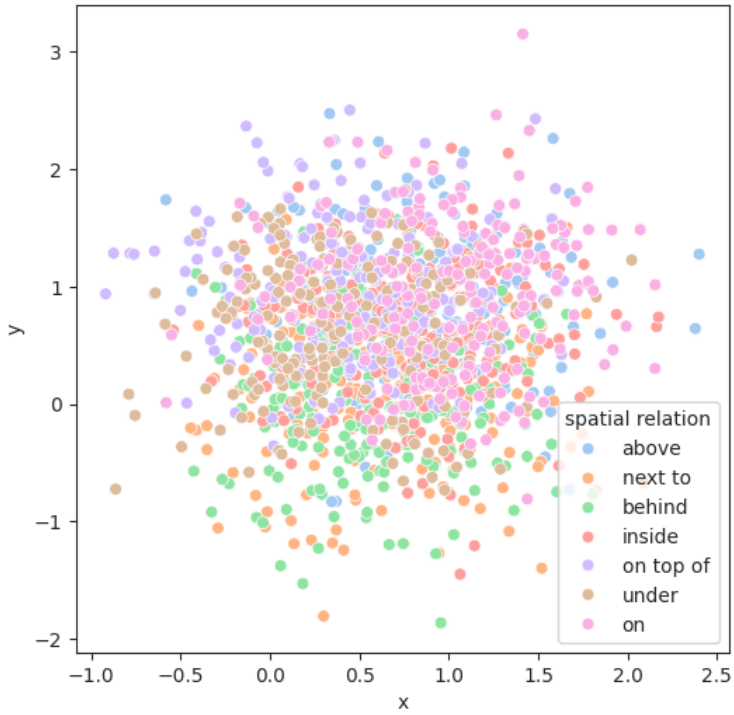

Figure 2: Scatter of 200 samples across each 7 common spatial relations overlapping severely to illustrate under-sensitivity on vision details. To begin, we perform average pooling on the visual tokens along the token length dimension. Next, we use t-SNE (t-Distributed Stochastic Neighbor Embedding) from sklearn to reduce the dimensionality of these pooled tokens. Finally, we plot the results as a scatter plot.

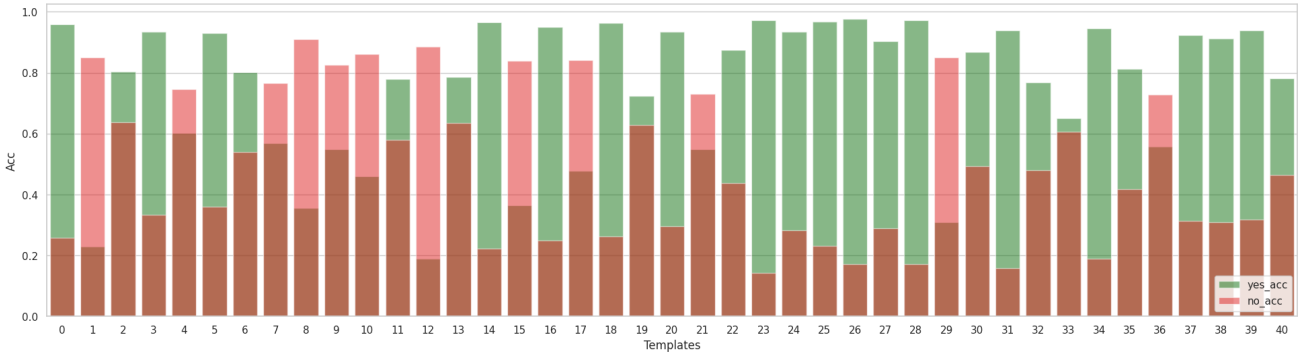

Figure 3: Results of comparison on "yes" and "no" question accuracy answered by LLaVA1.5 13B on VSR across 41 manually generated templates that got answered proportion over 0.85. The accuracy of the "yes" questions is plotted in the green bar and the "no" questions in the red bar.

of the second row and the missing character in the fourth column illustrate these issues. To address this, we will impose restrictions on using augmented images, employing them only during the pre-training phase. This will be discussed in detail in the section on constructing the training dataset.

2.2 Image-Text to Image

It is similar to text-to-image, but in addition to a prompt, an initial image is encoded to latent space then the noise is added to it. Then the model takes a prompt and the noisy latent image, predicts the noise, and removes the predicted noise from the initial latent image.

In Figure 5, we display some of the repaint results. As observed, the images repainted using the diffusion model generally show minimal differences from the original images, with only slight changes in details. However, the core spatial relationships between the primary entities in the images remain consistent, ensuring that the key visual information is preserved.

2.3 Image Inpainting

It replaces or edits specific areas of an image. This makes it a useful tool to replace an image area with something entirely new. It relies on a mask to determine which regions of an image to fill in; the area to inpaint is represented by white pixels and the area to keep is represented by black. The white pixels are filled in by the prompt.

We present a few examples of Image Inpainting in Figure 6. We selected an entity with a relatively small bounding box from the "subject" and "object" categories, masked its bounding box, and then randomly selected another entity from the MSCOCO classification (supplemented with additional entities) to inpaint it.

3 Multi-Vision Encoder

In this section, we discuss the details of our merged multi-vision encoder. With sufficient VSR text and image data support, we incorporated the most used visual backbone CLIP[1] with SigLIP[2], DINOv2[3] and SAM[4] to fully explore the potential of combining visual features. In Table 7, we show the specific version details of the used visual backbone, as well as the size of the features from the last hidden layer. The language-guided contrastive model CLIP and SigLIP benefit from the massive scale of noisy web image-text data. But self-supervised encoders DINOv2 and segmentation SAM detect more fine-grained visual details which may profit the spatial reasoning process. We then merge the pretrained backbones on the shelf to obtain a superior vision encoder focusing on visual spatial relations. Specifically, as shown in Figure 7 and Figure 8, we design porjectors and adapter to align visual tokens.

As summarized in Table 7, for CLIP, SigLIP, and DINOv2, we followed the approach used in LLaVA for aligning CLIP features. We utilized the features from the last hidden state layer as input and employed a projector to align them with the text tokens in Figure 7. Consistent with the LLaVA1.5's $mlp2x\_gelu$ projector design, a combination of fully connected layers and GELU activation layers is used to align the visual features. AS for the SAM, in Figure 8, we use two convolutional layers to process the SAM feature map, and then, we flatten it along the H and W dimensions into visual tokens.

Finally, we concatenate these tokens along the feature dimension following the method of MOF, MG-LLaVA and Cambrian.

4 Testing and Training Data

In this section, we illustrate details of our testing and training dataset construction. Table 8 shows the statistics of our training and testing data overall.

4.1 Testing Set

To maintain consistency with previous evaluations, we used the VSR zero-shot test sets with 1222 samples as the basis. We organized our test data into two sets:

(1)*Test-G* random sampled a prompt from the instruction fine-tune 50 templates pool for each triplet to evaluate instruction-following generalization ability during spatial reasoning. The instruction fine-tune 50 templates pool consists of the top 30 templates that performed best in the re-evaluation, as well as 20 additional templates generated by GPT-4. The GPT-4 generated templates are listed in Table 9.

(2)*Test-S* froze the template to the specific one

[1] https://huggingface.co/openai/clip-vit-large-patch14-336

[2] https://huggingface.co/google/siglip-so400m-patch14-384

[3] https://huggingface.co/facebook/dinov2-base

[4] https://huggingface.co/facebook/sam-vit-base

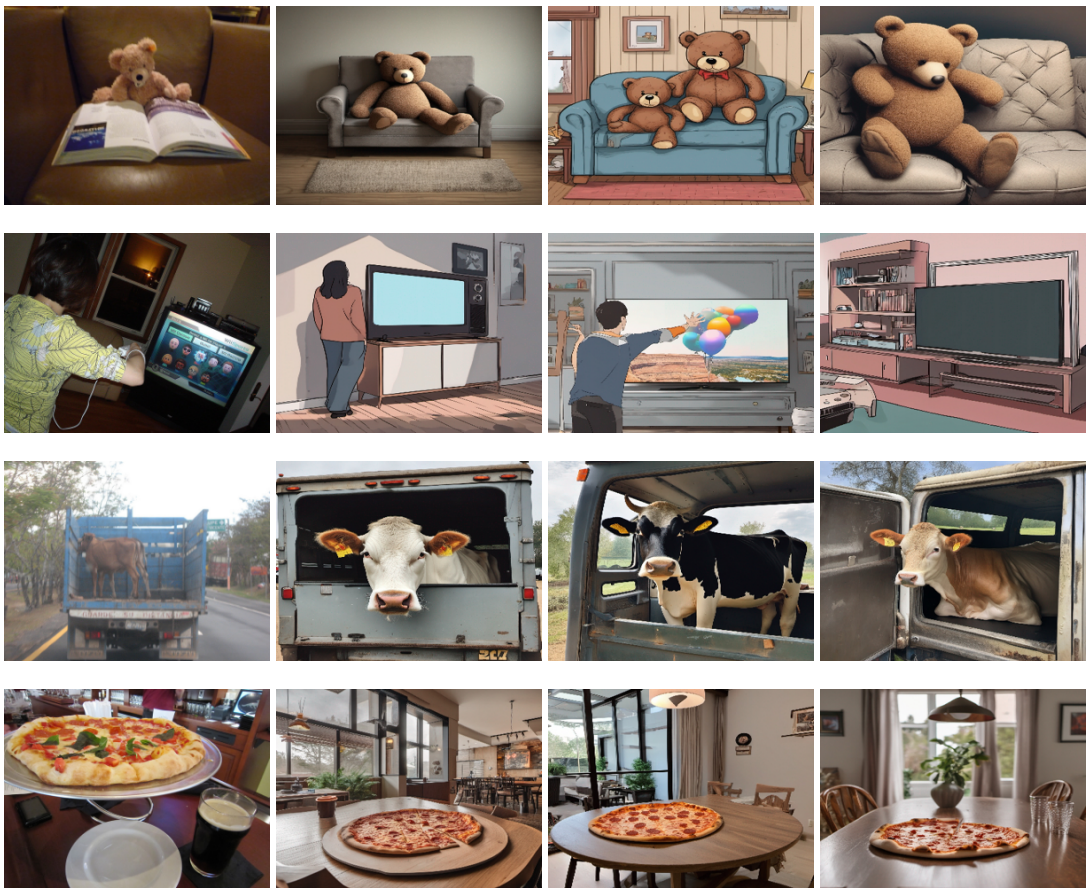

Figure 4: The *Test to image* exapmles of the repainting engine. The first column represents the original image. And the 3 columns left behind represent the *Test to image* painting result of the generated diffusion model. The control texts used are labeled from top to bottom as “The teddy bear is in the couch”, “The person is at the left of the TV”, “The cow is in the truck”, and “The pizza is on the table”.

Vision Backbone	Version	Selected Feature Size	Alignment Module
CLIP	openai/clip-vit-large-patch14-336	[1, 577, 1024]	projector
SigLIP	google/siglip-so400m-patch14-384	[1, 729, 1152]	projector
DINOv2	facebook/dinov2-base	[1, 257, 768]	projector
SAM	facebook/sam-vit-base	[1,64,64,256]	adaptor

Table 7: The specific version details of the used visual backbone, as well as the size of the features from the last hidden layer. Notably, LLaVA 1.5 uses the penultimate(-2) layer features from CLIP.

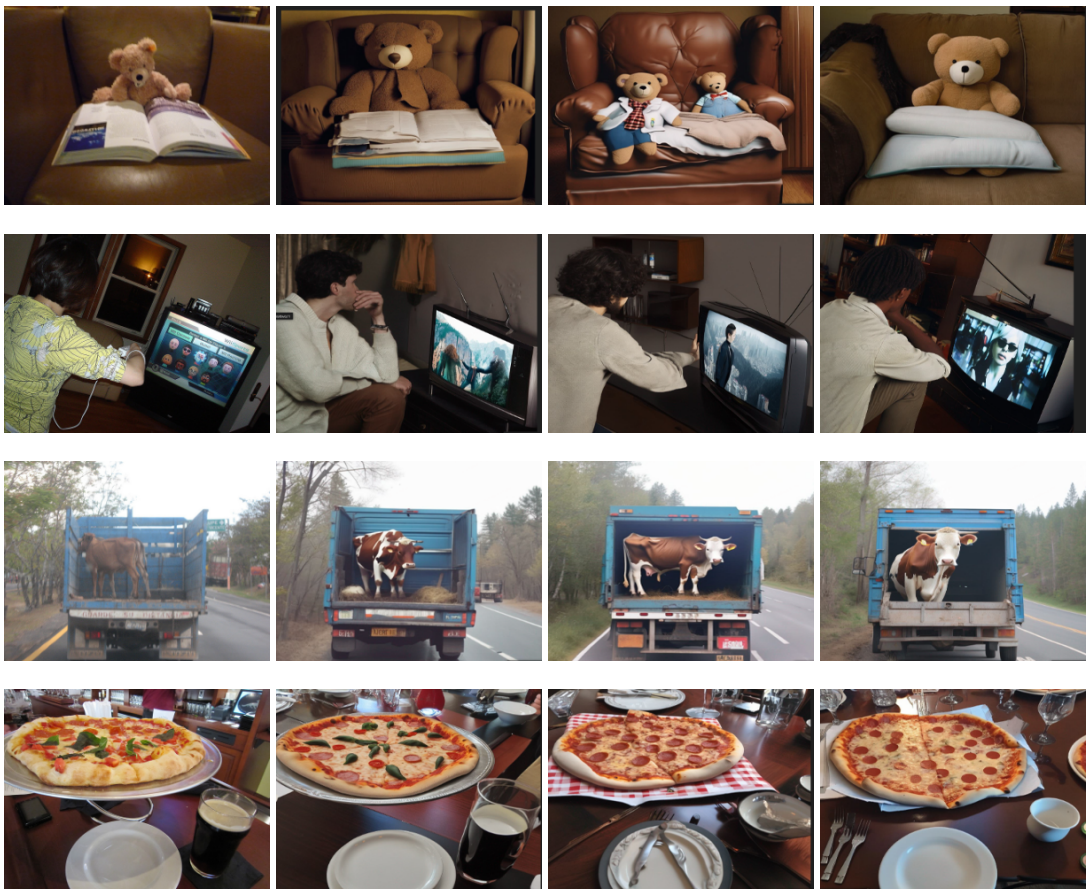

Figure 5: The *Image-Test to image* exapmles of the repainting engine. The first column represents the original image. And the 3 columns left behind represent the *Test to image* painting result of the generated diffusion model. The control texts used are labeled from top to bottom as “The teddy bear is in the couch”, “The person is at the left of the TV”, “The cow is in the truck”, and “The pizza is on the table”.

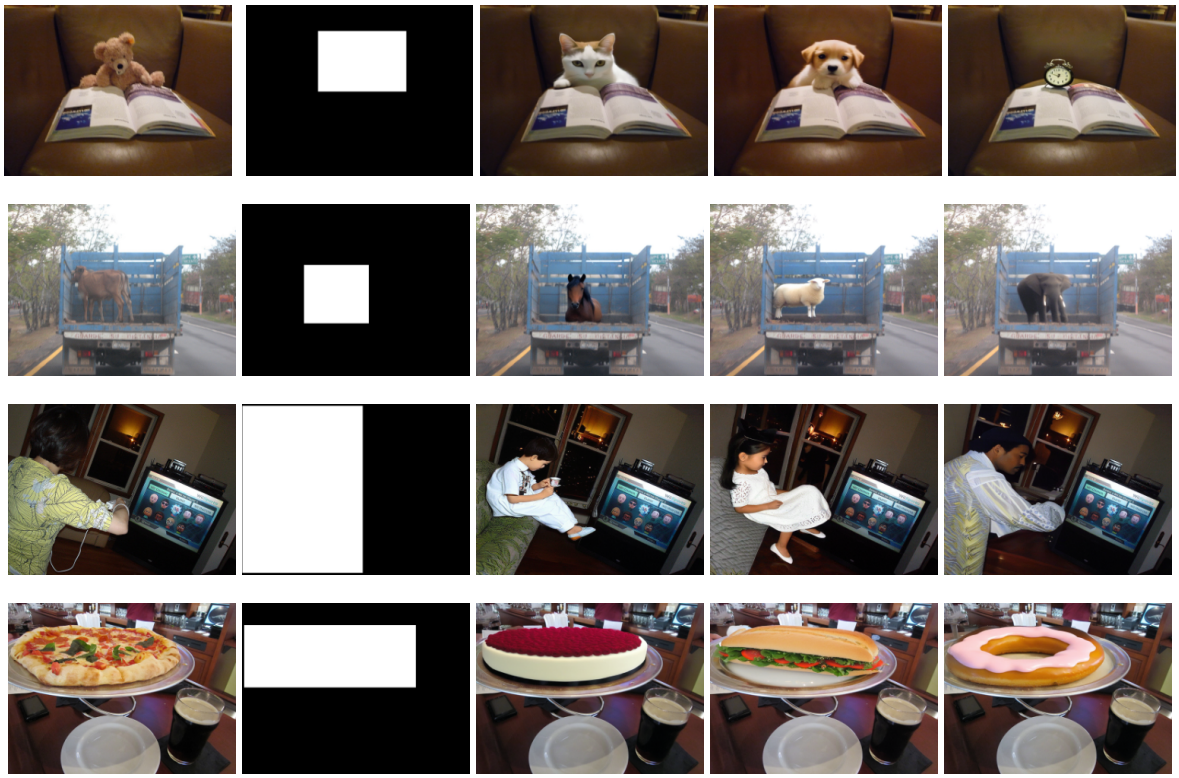

Figure 6: The *inpainting* exapmles of the repainting engine. The first column represent the original image. The second column represent the inpainting mask. And the 3 columns left behind represen the painting result of the generate diffusion model. For example we replace the teddy bear to cat, dog and clock in first row. In second row, we repaint the cow by horse, sheep and elephant. Then we displaced the person in the third row with girl, boy and man. And we substitute the pizza with cake, hot-dog and donut in the last row.

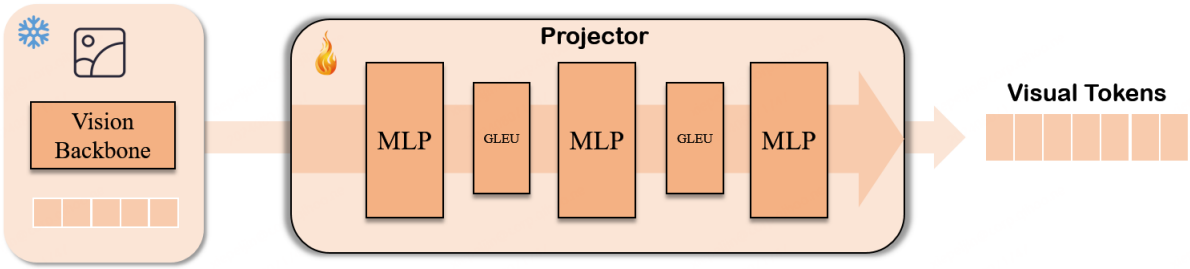

Figure 7: Details of the projector design. Consistent with the LLaVA 1.5 "$mlp2x\_gelu$" projector design, a combination of fully connected layers and GELU activation layers is used to align the visual features.

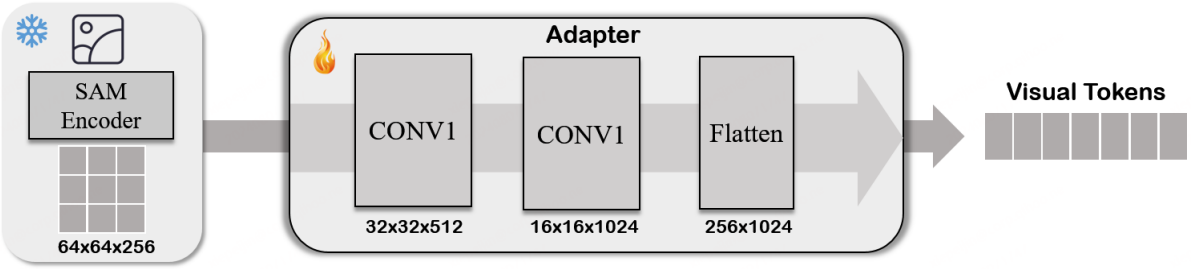

Figure 8: Details of the adapter design. We use two convolutional layers to process the SAM feature map, and finally, we flatten it along the H and W dimensions into visual tokens.

([*caption*], True or false.) which performs best and is the simplest during the re-evaluation process. The split aims to test the model's VSR capability on visual under its most proficient question-asking format and to avoid interference from other unrelated textual words.

4.2 Training Set

Excluding the test set, we collected 11k triplets with images from the original VSR dataset (including train and dev set) as a seed. Then we expand it several dozens of times into pre-train and instruction fine-tune data as follows:

Pre-training data: Under the three settings with a ratio of 5:3:2 (*image-text to image, text to image, image inpainting*), we repainted the original images. Furthermore, to prevent the model from being misled by occasional issues such as entity hallucinations and detail inconsistencies that can arise during the image generation process, we cautiously used the expanded image data only during the pre-training phase. This approach was specifically aimed at optimizing the projector and adapter components used for alignment.

Then we expand the quantity 20 to 100 times the original amount. Based on the original turning dataset, we first expanded the dataset in terms of the image amount. For instance, we took the original 5k images and applied a 20x repainting process to generate 100k vision samples. On the text side, then, we randomly selected templates from a pretrained pool of 10 GPT-4-generated templates, combining them with the triplet data to create the prompt inputs for pre-training. The pretrained GPT-4-generated templates pool is listed in the Table 10. We label the set as "pre-100k" for 100k pre-training data and "pre-500k" for 500k.

Instruction Fine-tuning data: We used general 50 prompt templates used in Test-G (30 manual and 20 GPT4-generated) to expand the 11k triplet data nearly 50 times to 500k, then name it as "turn-g 500k". Note that "turn-s 10k" and "turn-g 10k" are unexpanded instruction Fine-tuning data data for comparison. "turn-g 10k" used a randomly selected temple for each triplet and "turn-s 10k" used the specific template same as *Test-S*. As high-quality fine-tuning data, although limited in quantity and not sufficient for adjusting the LLM, our experiments have found that it can still lead to some effective improvements in the model's overall accuracy when used solely for optimizing the projector.

5 Average Intra-Class Distance

We use the metric Average Intra-Class Distance to represent the model's ability to summarize and generalize single positional concepts. The Average Intra-Class Distance is a metric used to evaluate the compactness of data points within the same class in a classification problem. It measures how close or far apart the data points within a single class are to each other. A lower average intra-class distance indicates that the points within a class are closely clustered, suggesting that the class is well-defined and distinct from other classes. Conversely, a higher average intra-class distance might indicate that the class is more spread out, which could suggest overlapping with other classes or less homogeneity within the class.

In this work, we first used Scikit-learn's dimensionality reduction tool t-distributed stochastic neighbor embedding (t-SNE) to reduce the dimensionality of average-pooled visual tokens to two dimensions. We then normalized the binary feature data of all categories, which included key statistical information, before calculating the intra-class distance and average intra-class distance by euclidean distance.

The euclidean distance between two points (x_i, y_i) and (x_j, y_j) is given by:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

For each class k, the intra-class average distance D_k is computed as:

$$D_k = \frac{2}{n_k(n_k - 1)} \sum_{i=1}^{n_k - 1} \sum_{j=i+1}^{n_k} d_{ij}$$

where n_k is the number of samples in class k.

The overall intra-class average distance across all classes is:

$$D_{\text{avg}} = \frac{1}{K} \sum_{k=1}^{K} D_k$$

where $K = 7$ is the number of classes.

To summarize, we used Intra-Class Distance to evaluate the model's understanding and summarization of the extracted visual tokens for each positional relationship category. A smaller value indicates better clustering within that category, suggesting a deeper and more accurate understanding by the model. Along with the scatter plots, only the processed visual features exhibit relatively

Data	Usage	Repaint	Rewrite	Amount	Triplet	Images	Templates
Test-G	test	-	-	1222	1222	715	30+20
Test-S	test	-	-	1222	1222	715	1
turn-s 11k	PT / IFT	×	×	3489+7680	3489+7680	5544	1
turn-g 11k	PT / IFT	×	✓	3489+7680	3489+7680	5544	30+20
turn-g 500k	IFT	×	✓	500k	3489+7680	5544	30+20
pre-100k	PT	✓	✓	100k	10k	100k	10
pre-200k	PT	✓	✓	200k	10k	200k	10
pre-300k	PT	✓	✓	300k	10k	300k	10
pre-400k	PT	✓	✓	400k	10k	400k	10
pre-500k	PT	✓	✓	500k	10k	500k	10

Table 8: Statistics of training and testing data. PT refers to pertaining and IFT to instruction fine-tuning. The "Repaint" indicates whether the training data used augmented repainting image data and the "Rewrite" indicates whether used augmented rewriting text data from the template pool. "Template" calculates the number of templates used in the instruction construction.

small Intra-Class Distances and larger inter-class distances, do we believe that such visual features are sensitive to positional information, which is crucial for enhancing the VLLM's ability to perform visual spatial reasoning.

References

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *Preprint*, arXiv:2308.01390.

Gorka Azkune, Ander Salaberria, and Eneko Agirre. 2024. Grounding spatial relations in text-only language models. *Neural Networks*, 170:215–226.

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

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *Preprint*, arXiv:2310.09478.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Imagebind-llm: Multi-modality instruction tuning. *Preprint*, arXiv:2309.03905.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2023. Bliva: A simple multimodal llm for better handling of text-rich visual questions.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *Preprint*, arXiv:2102.03334.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. 2024. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *Preprint*, arXiv:2402.12058.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *Preprint*, arXiv:2305.03726.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *Preprint*, arXiv:1908.03557.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *Preprint*, arXiv:2311.07575.

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

Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. 2024. Prismer: A vision-language model with multi-task experts. *Preprint*, arXiv:2303.02506.

Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, and Ping Luo. 2023a. Tiny lvlm-ehub: Early multimodal experiments with bard.

Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, and Ping Luo. 2023b. Tiny lvlm-ehub: Early multimodal experiments with bard. *Preprint*, arXiv:2308.03729.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *Preprint*, arXiv:1908.07490.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl: Modularization empowers large language models with multimodality. *Preprint*, arXiv:2304.14178.

Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. 2023a. Vpgtrans: Transfer visual prompt generator across llms. *Preprint*, arXiv:2305.01278.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *Preprint*, arXiv:2303.16199.

Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. 2024. Cobra: Extending mamba to multi-modal large language model for efficient inference. *Preprint*, arXiv:2403.14520.

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

Dongsheng Zhu, Xunzhu Tang, Weidong Han, Jinghui Lu, Yukun Zhao, Guoliang Xing, Junfeng Wang, and Dawei Yin. 2024. Vislinginstruct: Elevating zero-shot learning in multi-modal language models with autonomous instruction optimization. *Preprint*, arXiv:2402.07398.

Index	Prompt Template
1	In the image provided, does the relationship between [subject] and [object] match the relationship described as [relation] in the triple?
2	Based on the visual content, is the [relation] between [subject] and [object] accurately depicted in the image?
3	Is the connection between [subject] and [object] in the image reflective of the [relation] stated in the triple?
4	Does the image illustrate a [relation] relationship between [subject] and [object] as described in the triple?
5	Can the relationship [relation] between [subject] and [object] be confirmed from the information shown in the image?
6	Is the relationship [relation] between [subject] and [object] evident from the image content?
7	In the image, does the interaction between [subject] and [object] support the [relation] relationship given in the triple?
8	Does the image visually represent [relation] as the relationship between [subject] and [object]?
9	Is the [relation] between [subject] and [object] consistent with what is depicted in the image?
10	Can you confirm that the image portrays a [relation] between [subject] and [object], as described in the triple?
11	Does the visual interaction between [subject] and [object] in the image align with the [relation] described?
12	Is the image consistent with the idea that [subject] and [object] share a [relation] relationship?
13	Does the image support the assertion that [subject] has a [relation] with [object]?
14	In the image, does the [relation] relationship between [subject] and [object] hold true?
15	Does the image depict a scenario where [subject] and [object] are in a [relation] relationship?
16	Is the [relation] between [subject] and [object] in the triple reflected in the visual content of the image?
17	Does the image show [subject] and [object] in a way that confirms a [relation] relationship?
18	Is the relationship [relation] between [subject] and [object] visually apparent in the image?
19	Does the image verify the [relation] relationship between [subject] and [object] mentioned in the triple?
20	Is the relationship [relation] depicted between [subject] and [object] consistent with the image?

Table 9: 20 Prompt Templates that generated by GPT-4 used for Test-G and turn-g 500k.

Index	Prompt Template
1	Describe the spatial relationship between [subject] and [object] in the image.
2	Identify and explain the positional connection between [subject] and [object] as shown in the image.
3	Detail the spatial interaction between [subject] and [object] in the image.
4	Provide a description of how [subject] and [object] are related in the image.
5	Analyze and describe the spatial relationship between [subject] and [object] based on the image.
6	Explain the nature of the positional relationship between [subject] and [object] in the image.
7	Articulate the relationship between [subject] and [object] as depicted in the image.
8	Clarify the positional connection between [subject] and [object] in the image.
9	Describe how [subject] and [object] are interacting in the image.
10	Give a detailed account of the spatial relationship between [subject] and [object] shown in the image.

Table 10: Prompt templates to command model to describe spatial relationship in image during the pretrained data construction.