VolDoGer: LLM-assisted Datasets for Domain Generalization in Vision-Language Tasks

Anonymous ACL submission

Abstract

Domain generalizability is a crucial aspect of a deep learning model since it determines the capability of the model to perform well on data from unseen domains. However, research on the domain generalizability of deep learning models for vision-language tasks remains limited, primarily because of the lack of required datasets. To address these challenges, we propose VolDoGer: **V**isi**o**n-**L**anguage Dataset for **Do**main **Ge**ne**r**alization, a dedicated dataset designed for domain generalization that addresses three vision-language tasks: image captioning, visual question answering, and visual entailment. We constructed VolDoGer by extending LLM-based data annotation techniques to vision-language tasks, thereby alleviating the burden of recruiting human annotators. We evaluated the domain generalizability of various models through VolDoGer.

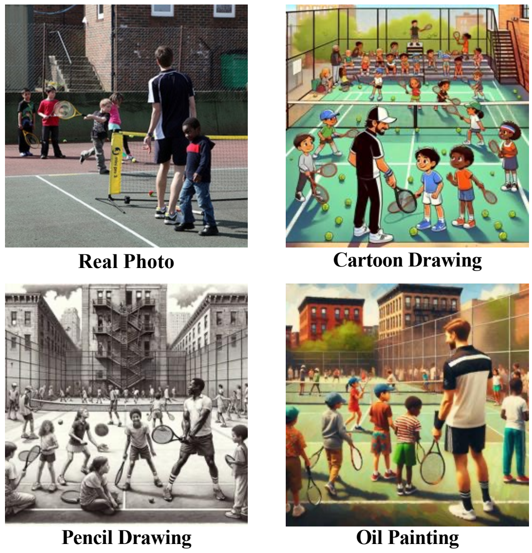


Figure 1: Examples of images with various styles in VolDoGer. Please refer to Appendix F for more examples with annotation.

1 Introduction

Vision-language models have evolved and demonstrated outstanding performance in various tasks (Chen et al., 2023) such as image captioning (Stefanini et al., 2022), visual question answering (VQA) (Wu et al., 2017; de Faria et al., 2023), and visual entailment (VE) (Xie et al., 2019). However, these vision-language models can suffer from domain shift, which is a significant challenge for deep learning models (Wang and Deng, 2018; Fang et al., 2024). Domain shift refers to a phenomenon in which the domain of the data changes between the training and inference phases of a model. For example, an image classification model trained on photos may not perform well when applied to sketch images (Zhou et al., 2022). This issue is prevalent in NLP tasks (Elsahar and Gallé, 2019; Ramponi and Plank, 2020; Calderon et al., 2023) to vision-language tasks (Chen et al., 2017; Zhao et al., 2017; Yang et al., 2018; Zhao et al., 2020).

Extensive research has been conducted on domain generalization to mitigate domain shift (Zhou et al., 2022; Wang et al., 2022). These lines of study aim to utilize multiple source domains to enhance the generalizability of the model against out-of-domain target domains. However, the difficulty of collecting annotated data from various source domains may diminish the practicality of domain generalization. Although it is relatively simple to gather data for unimodal tasks such as image classification or text classification (Blitzer et al., 2007; Peng et al., 2019), it may be more difficult to collect data for multimodal tasks because they require a *pair* of data in each modality.

Consequently, there is a lack of datasets for domain generalization in multimodal tasks, including vision-language tasks. The absence of a dedicated dataset makes it difficult to explore domain generalization in vision-language tasks. For example, existing studies on domain generalization for image captioning merged multiple existing datasets with different subjects, each of which contains real photos as input images (Ren et al., 2023). However, this approach fails to fully consider the diversity of domains because it contains only real photographs, thereby not accounting for the domain shift in the *style* of the input image. Furthermore, recent advancements in generative models have led to a significant increase in the volume of generated content encompassing a diverse array of styles. In view of this challenge, vision-language models must be capable of delivering accurate and consistent results on generated images with various styles as well, considering that generative models can easily produce images in designated styles (Zhang et al., 2023a). Hence, a specialized dataset for domain generalization in vision-language tasks is required to address these challenges and ensure a robust performance across diverse image styles.

However, it is difficult to construct such a dataset through the collection and annotation by human annotators. Unlike relatively straightforward tasks such as image classification, where images with various styles can be collected with a simple search (e.g., “aeroplane painting”) (Peng et al., 2019), creating a vision-language task dataset for domain generalization imposes more severe restrictions. For instance, a dataset for domain generalization in image captioning tasks would require a large set of similar images in different styles, such as cartoons, paintings, and sketches, as well as their descriptions. Moreover, these tasks require more complex human annotation procedures than simple tasks, leading to higher annotation costs and more efforts for quality control (Rashtchian et al., 2010).

To address these challenges and effectively construct datasets for domain generalization in vision-language tasks, we propose leveraging large language model (LLM)-based data annotation (Tan et al., 2024). LLM-based data annotation uses LLMs as data annotators to replace human annotators. Researchers have found this strategy to be cost-effective in producing consistent results compared with human annotators (Wang et al., 2021; Ding et al., 2023). However, previous studies on LLM-based data annotation have primarily focused on text data (Li et al., 2023b; Zhang et al., 2023b; He et al., 2024; Bansal and Sharma, 2023). Although recent studies have applied LLM-based data annotation to image captioning tasks, they have not considered image data and relied solely on text input (Choi et al., 2024). In this study, we leverage recent advancements in LLM with improved image interpretation capabilities, such as GPT-4o (OpenAI, 2023, 2024), and explore the use of LLMs as multimodal data annotators by collaborating with recent image generation models (Betker et al., 2023; Esser et al., 2024).

Using the proposed multimodal LLM-based data annotation, we constructed VOLDOGER: **V**isi**o**n-**L**anguage Dataset for **Do**main **Ge**ne**r**alization, which is the first dedicated dataset designed to facilitate domain generalization across three vision-language tasks: image captioning, VQA, and VE. VOLDOGER involves four different styles, which are real photos, cartoon drawings, pencil drawings, and oil paintings. Figure 1 showcases an example of image with various styles consisting VOLDOGER. Based on these source data encompassing various styles, it is possible to train a model with improved domain generalizability using VOLDOGER. In this study, we utilized VOLDOGER to validate the presence of domain shifts in these tasks and to evaluate the effectiveness of existing domain generalization techniques.

Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first work to establish multimodal LLM-based data annotation while considering multimodal inputs.
- We release VOLDOGER, a first dedicated dataset designed to advance research on domain generalization across three vision-language tasks.
- Our extensive experiments demonstrate the presence of domain shift and the effectiveness of domain generalization techniques in vision-language tasks.

2 Related Work

2.1 Domain Generalization for Vision-Language Tasks

Despite the lack of a dedicated dataset for domain generalization in vision-language tasks, researchers are increasingly exploring this area. For example, a relevant study proposed a framework for domain generalization in image captioning (Ren et al., 2023). They incorporated the use of text data through visual word guidance and sentence similarity based on previous research (Wang et al., 2020). However, although they proposed an effective framework, the datasets they used, such as MSCOCO (Chen et al., 2015) and Flickr30k (Young et al., 2014) exhibited significant overlap. Furthermore, these datasets exhibit limited differences in visual features because they primarily consist of real photos. In contrast, our objective is to create datasets for various vision-language tasks, such as image captioning, that encompass diverse visual styles within images.

In the field of VQA, VQA-GEN (Unni et al., 2023) suggested constructing a dataset for domain generalization by modifying an existing dataset, which aligns with the purpose of our work. However, their manipulation strategies for visual features mostly consist of simple noise-based modifications, such as injecting blurs. Moreover, VQA-GEN is not publicly available, which reduces the usability of this study for future work. Another line of research proposed a methodology that enables task generalization on VQA datasets that require image understanding and compositional reasoning (Shrestha et al., 2019; Gamage and Hong, 2021).

To the best of our knowledge, this is the first study to explore domain generalization in visual entailment. Moreover, we propose a multimodal LLM-based data annotation pipeline and introduce VOLDOGER, which is a publicly available dataset constructed using our pipeline, to facilitate future advancements in domain generalization for vision-language tasks.

2.2 LLM-based Data Annotation

As LLMs exhibit various capabilities, researchers have explored leveraging them as data annotators to replace human annotators. For example, automatic annotation through GPT-3 has demonstrated superior downstream model performance compared with human performance at a lower cost (Wang et al., 2021; Gilardi et al., 2023). Furthermore, the capability of GPT-3 to generate labeled data from scratch was demonstrated (Ding et al., 2023). Consequently, the exploration of LLM applications as data annotators continues to expand, underscoring their utility in streamlining and optimizing the data annotation process (Li et al., 2023b; Zhang et al., 2023b; He et al., 2024; Bansal and Sharma, 2023).

A recent study closely related to our work also suggested data annotation for image captioning tasks, one of the tasks that we address (Choi et al., 2024). However, they made limited use of multimodal data since they did not consider image inputs while annotating the given data. Instead, they only paraphrase the given text input and translate the paraphrases into another language. By contrast, we aim to actively utilize LLMs as multimodal data annotators.

3 LLM-based Data Annotation for Vision-Language Tasks

In this section, we introduce the proposed framework for multimodal LLM-based data annotation for vision-language tasks, as illustrated in Figure 2. The framework comprises two primary phases: stylized image generation and label annotation. Although the stylized image generation process was shared across the three tasks, the label annotation process varied slightly to accommodate the specific characteristics of each task. The objective of this framework is to convert the given dataset $\mathcal{D}_{ori}$ into a transferred dataset with the designated style $\mathcal{D}_{sty}$, where the input image x_{ori} is transformed into stylized image x_{sty}. We utilize a multimodal LLM $\mathcal{M}$ to perform this transformation. For the exact instruction prompts, please refer to Appendix G.

3.1 Stylized Image Generation

In the first phase, the framework aims to create an image x_{sty} that retains the content and semantics of x_{ori} but has a designated style. This phase consisted of four steps: image decomposition, style injection, image generation, and image verification.

Image Decomposition. We first input x_{ori} from $\mathcal{D}_{ori}$ into $\mathcal{M}$ with the instruction P_{ID} to generate a prompt describing semantics in x_{ori}, $p_{ori} = \mathcal{M}(P_{ID}, x_{ori})$, which can be used to reconstruct x_{ori} through an image generation model $\mathcal{G}$.

Style Injection. Next, we transform p_{ori} into a stylized prompt $p_{sty} = \mathcal{M}(P_{SI}^{sty}, p_{ori})$ based on instruction P_{SI}. The generated p_{sty} retains the content of x_{ori} while incorporating information about the desired style.

Image Generation. In this step, we pass the stylized prompt p_{sty} to the text-to-image generation model $\mathcal{G}$. Subsequently, a transformed image with the desired style x_{sty} is generated by $x_{sty} = \mathcal{G}(p_{sty})$. Appendix F.1 provides the generated image x_{sty} and its prompt p_{sty}.

Image Verification. It is important to note that the generated x_{sty} may not fully capture the core semantics of x_{ori}. The distinction between x_{ori} and x_{sty} could complicate the subsequent annotation process; that is, the original label may not correspond to x_{sty} if it deviates significantly from x_{ori}.

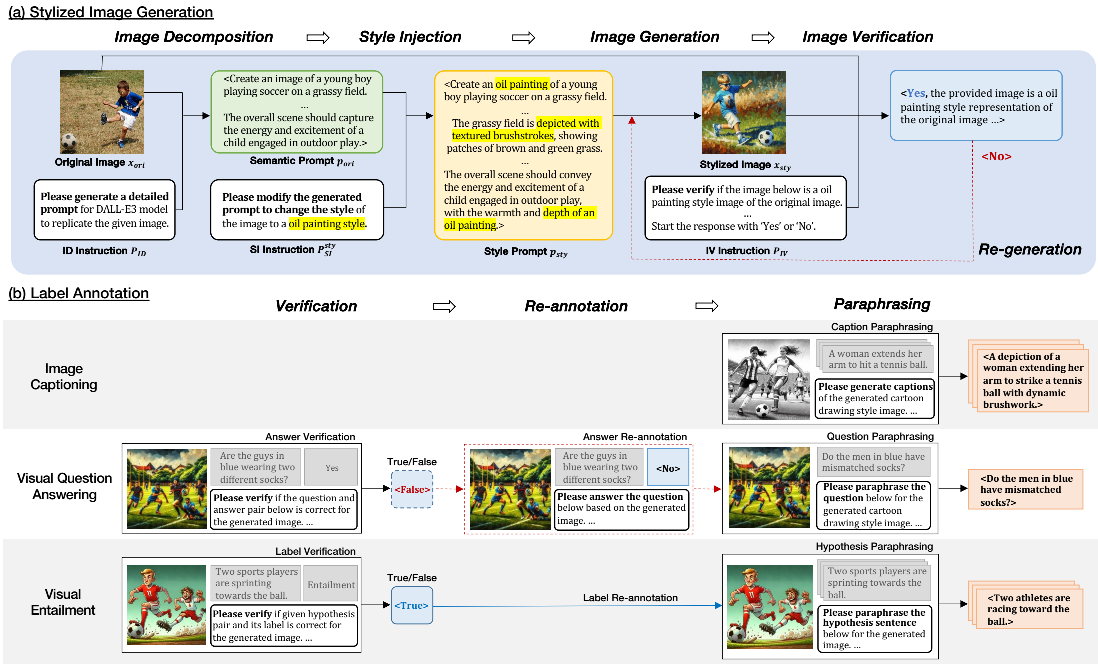


Figure 2: Overall procedure for annotating data through our proposed framework. (a) illustrates the process for stylized image generation in Section 3.1, and (b) displays the label annotation process for each task, which is described in Section 3.2, 3.3, and 3.4.

To address this issue, we propose introducing an image verification step for the generated x_{sty}. We simultaneously pass x_{sty} and x_{ori} to $\mathcal{M}$ and compare the two images to verify whether x_{sty} retains the essential content of x_{ori}. This is formulated as $v_{IV} = \mathcal{M}(P_{IV}, x_{ori}, x_{sty})$. If the candidate x_{sty} does not pass the verification; in other words, if v_{IV} is false, we return to the image generation step and create a new x_{sty}. This verification process is crucial for ensuring the accuracy and consistency of x_{sty} relative to x_{ori}, thereby maintaining the validity of the annotations.

After completing the verification process, we proceed to the next phase: label annotation.

3.2 Label Annotation for Image Captioning Task

The image captioning task aims to generate a description y from a given image x. Unlike the other two tasks, the image captioning task does not require any additional input besides the image. Our goal is to create a data pair $d_{sty} = (x_{sty}, y_{sty})$ for the task.

Caption Paraphrasing. Instead of directly assigning the original y_{ori} to the generated x_{sty}, we generate a paraphrase of y_{ori} as y_{sty}, while considering the style of x_{sty}. This process is beneficial for offering the model diverse expressions (Fan et al., 2023). It is also crucial to avoid duplicating the label data, which can negatively impact the training procedure (Schofield et al., 2017). To this end, we pass x_{sty} and y_{ori} to $\mathcal{M}$, obtaining $y_{sty} = \mathcal{M}(P_{CP}, x_{sty}, y_{ori})$.

3.3 Label Annotation for Visual Question Answering Task

The purpose of VQA task is to answer question q based on the given image x. The VQA model takes x and q as inputs and predicts the answer y. We aim to create a data pair for the VQA task as $d_{sty} = (x_{sty}, q_{sty}, y_{sty})$.

Answer Verification. Although x_{sty} may pass the image verification step, the original label y_{ori} for question q_{ori} may not be valid for x_{sty} owing to minor differences. For instance, if the question q_{ori} was "How many cups are on the table?" and x_{ori} had two cups, but x_{sty} contains four cups, the original label y_{ori} "two" would no longer be valid for x_{sty}.

To verify y_{ori} with respect to x_{sty}, we utilize $\mathcal{M}$ to confirm if y_{ori} is the answer to q_{ori} given x_{sty} based on $v_{AV} = \mathcal{M}(P_{AV}, x_{sty}, q_{ori}, y_{ori})$. If y_{ori} is a valid answer to q_{ori} given x_{sty}, we assign y_{ori} as y_{sty}. Otherwise, if y_{ori} is not a valid answer for q_{ori} given x_{sty}, we proceed to the answer re-annotation step, as detailed below.

Answer Re-annotation. In cases where y_{ori} is in-correct for q_{ori} given x_{sty}, We simply employ $\mathcal{M}$ to answer q_{ori} and generate $y_{sty} = \mathcal{M}(P_{AR}, x_{sty}, q_{ori})$.

Question Paraphrasing. Similar to the image captioning task, we paraphrase the given q_{ori} to address the issue of duplication. This step is more crucial in VQA tasks because the allocation of identical question phrases and answers between different images can induce shortcut learning to focus solely on the question (Ramakrishnan et al., 2018; Agrawal et al., 2018; Jing et al., 2020; Guo et al., 2021). To address this concern, we obtain q_{sty}, a paraphrased version of q_{ori}, using $q_{sty} = \mathcal{M}(P_{QP}, q_{ori})$.

3.4 Label Annotation for Visual Entailment Task

The visual entailment task is similar to the natural language inference task. Instead of predicting the entailment of a text premise and hypothesis, the visual entailment task involves taking an image as a premise and predicting the entailment of the premise and text hypothesis. In this task, we create a data pair $d_{sty} = (x_{sty}, h_{sty}, y_{sty})$, where h_{sty} represents the hypothesis.

Label Verification. It is important to ensure the validity of label y_{ori} in relation to the newly generated x_{sty}. To accomplish this, we use $\mathcal{M}$ to verify y_{ori} given (x_{sty}, h_{ori}), acquiring $v_{LV} = \mathcal{M}(P_{LV}, x_{sty}, h_{ori}, y_{ori})$. If y_{ori} is not the correct label for x_{sty} and h_{ori}, we proceed to label the re-annotation step as described below. Otherwise, we assign y_{ori} as y_{sty}.

Label Re-annotation. If y_{ori} is not valid for h_{ori} given x_{sty}, we utilize $\mathcal{M}$ to obtain $y_{sty} = \mathcal{M}(P_{LR}, x_{sty}, h_{ori})$.

Hypothesis Paraphrasing. Similar to VQA task, the use of identical hypotheses can lead to shortcut learning (Geirhos et al., 2020). To address this issue, we assign the paraphrase of h_{ori} as the hypothesis h_{sty} for x_{sty}. We guide $\mathcal{M}$ to paraphrase h_{ori} and obtain $h_{sty} = \mathcal{M}(P_{HP}, h_{ori})$.

4 VolDoGer

Based on the annotation framework discussed in the previous section, we constructed VolDoGer, a dedicated dataset for domain generalization for vision-language tasks. In this section, we introduce VolDoGer and detail the data configuration and statistics of each dataset. In addition to the realistic photos from the original datasets, VolDoGer includes four distinct image styles: real photos[1], cartoon drawings, pencil drawings, and oil paintings. For more detailed information and analyses, please refer to Appendix C.

	R	C	P	O	
R	-	0.0194	0.0244	0.0303	
C	0.0128	-	0.0121	0.0134	**Average**
P	0.0175	0.0117	-	0.0164	0.0193
O	0.0127	0.0096	0.0111	-	0.0126

Table 1: Domain gap of each style in VolDoGer-Cap, measured with MMD by ResNet and BERT output vectors. Orange figures denote the visual domain gap, and blue figures represent the linguistic domain gap.

4.1 VolDoGer-Cap

VolDoGer-Cap is a part of VolDoGer designed for image captioning tasks. To construct this dataset, we utilized the UIT-VIIC dataset (Lam et al., 2020), which is a subset of the MSCOCO captioning dataset (Chen et al., 2015) focused on sports images, following a previous study (Choi et al., 2024). Consequently, VolDoGer-Cap contains 2695 images for training, 924 images for validation, and 231 images for testing each style. Each image is associated with five different captions.

To identify the domain gap of each style in VolDoGer-Cap, we use the maximum mean discrepancy (MMD) (Gretton et al., 2006) to measure the difference in the visual and linguistic features of each domain, following previous studies (Zhang et al., 2021; Chen et al., 2021; Ren et al., 2023). Specifically, we leveraged the encoded vectors of ResNet (He et al., 2016) and BERT (Devlin et al., 2019) to extract features from the domain and computed the MMD distances using these features. Table 1 demonstrates the result of the analysis. In this analysis, we found that VolDoGer-Cap exhibited a remarkable visual domain gap across every domain compared with the collection of datasets based on real photos, which was adopted by previous study (Ren et al., 2023), revealing the value of VolDoGer for domain generalization in vision-language tasks.

4.2 VolDoGer-VQA

VolDoGer-VQA is built upon the question and answer from VQA-v2 (Goyal et al., 2017), which utilizes the same images as the MSCOCO dataset

[1]Note that real photos indicate the original images taken from a camera with human-annotated data.

and UIT-VIIC. To enhance the efficiency of the data annotation, we extracted images from the UIT-VIIC dataset along with their corresponding questions and answers. To ensure the quality and consistency of LLM-based data annotation, we exclusively used yes/no questions, as they are less ambiguous and require more direct answers than open-ended or multiple-choice questions, which can vary significantly in complexity and interpretation. Consequently, this dataset comprises 2091 images with 4120 questions for training, 711 images with 1452 questions for validation, and 182 images with 340 questions for testing for each style.

4.3 VOLDOGER-VE

For VOLDOGER-VE, we used the SNLI-VE (Xie et al., 2019) dataset, which served as the primary dataset for the visual entailment task. Similar to the approaches for VOLDOGER-CAP and VOLDOGER-VQA, we used only images related to football. To achieve this, we selected images containing text premise that includes the words "soccer" or "football." Subsequently, we divided them into training, validation, and test sets in a ratio of 8:1:1. As a result, VOLDOGER-VE comprises 619 images with 7319 hypotheses for training, 77 images with 957 hypotheses for validation, and 78 images with 856 hypotheses for testing each style.

5 Experiment

In this section, we describe the extensive experiments conducted using our constructed VOLDOGER dataset and present several insights derived from the experimental results.

5.1 Experimental Setup

First, we briefly introduce the experimental setup used in our experiments. Various models are trained using different backbones for each task. For the domain shift experiment in Section 5.2, we fine-tuned the models with ViT (Dosovitskiy et al., 2021) and CLIP (Radford et al., 2021) encoders with a GPT-2 (Radford et al., 2019) decoder, as well as the BLIP (Li et al., 2022) model on VOLDOGER-CAP for the image captioning task. For the VQA and VE tasks, we trained the models using the ViT and CLIP image encoders with the BERT (Devlin et al., 2019) text encoder as well as the BLIP model on VOLDOGER-VQA and VOLDOGER-VE. Similarly, for the domain generalization experiments in Section 5.3, we trained the models using the ViT

Captioning ViT	Trained on Real Photos				
	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	44.71	50.54	27.99	0.6855	-4.6252
Cartoon	21.29	33.90	15.10	0.6167	-4.8264
Pencil	16.76	30.82	13.52	0.5948	-4.6103
Oil	13.02	26.91	12.10	0.5820	-4.7255
CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	41.29	47.76	25.98	0.6768	-4.6252
Cartoon	17.48	30.21	12.73	0.6021	-4.8315
Pencil	14.07	27.35	11.15	0.5742	-4.6123
Oil	11.14	24.31	10.57	0.5683	-4.7301
BLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	48.19	49.95	30.85	0.6950	-4.6671
Cartoon	22.65	32.05	16.12	0.6269	-4.9008
Pencil	17.52	28.45	14.16	0.5993	-4.7417
Oil	14.86	27.98	12.79	0.5868	-4.9053

Table 2: Experimental result demonstrating domain shift in image captioning tasks. In-domain performance is indicated in green, and out-domain performance is indicated in red. Please refer to Table 17 for the results of the models trained on other domains.

VQA	Trained on Real Photos			
	Real	Cartoon	Pencil	Oil
ViT	55.03	48.52	47.76	48.65
CLIP	58.23	49.11	50.41	49.41
BLIP	59.19	50.29	51.32	50.88
VE	**Trained on Real Photos**			
	Real	Cartoon	Pencil	Oil
ViT	72.15	52.51	57.72	58.78
CLIP	73.10	55.85	61.61	60.92
BLIP	66.73	48.26	52.93	53.62

Table 3: Experimental result demonstrating domain shift in VQA and VE tasks. In-domain performance is indicated in green, and out-domain performance is indicated in red. Please refer to Table 19 and 20 for the results of the models trained on other domains.

and frozen CLIP encoder models. In addition, we included a baseline that leveraged the ViT encoder with a dedicated technique for domain generalization, extending this approach to VQA and VE tasks (Ren et al., 2023).

To measure the performance of the model in the image captioning task, we employed various metrics such as BLEU (Papineni et al., 2002), ROUGE_L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2020), and BARTScore (Yuan et al., 2021), following previous study (Choi et al., 2024). The accuracy was used as a metric for the VQA and VE tasks. We report the average performance of the models trained with five random seeds. For more detailed information about the implementation of the experiment, please refer to Appendix B.

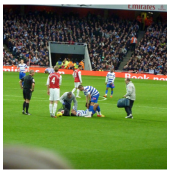
In-domain Data	Out-domain Data
 • A group of men playing a game of **soccer**.	 • A group of people standing on **top of a building**.
 • A group of young **children** playing a game of **soccer**.	 • A group of young **men** playing a game of **baseball**.
 • A group of young men playing a game of **soccer**.	 • A group of young men playing a game of **frisbee**.

Table 4: Examples of domain shift on image captioning task. The captions are produced by a ViT-based model trained on real photos. The left side of the table showcases in-domain examples, and the right side of the table showcases out-domain examples.

5.2 Existence of Domain Shift in Vision-Language Tasks

First, we investigate the existence of a domain shift using VOLDOGER. To accomplish this, we train each model on a single domain and test it across four domains: **R**eal photo, **C**artoon drawing, **P**encil drawing, and **O**il painting.

Tables 2 and 3 list the experimental results for the three tasks. In these experiments, we observed significant differences between the in-domain and out-domain performances, confirming the existence of a domain shift in response to input images with different styles. Examples of the outputs produced by a captioning model solely using real photos in Table 4 support the experimental results. In these examples, we can observe that the model cannot accurately generate descriptions for images with similar content but different styles. While this phenomenon has been observed in other tasks, such as image classification (Peng et al., 2019), validating its existence in vision-language tasks has been challenging because of the absence of a dedicated dataset. Our study demonstrates that this phenomenon persists in vision-language tasks us-

Captioning	**Trained on R+C+P**				
ViT	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	46.46	51.99	28.79	0.6849	-4.6446
Cartoon	42.83	42.07	23.66	0.6742	-4.8284
Pencil	42.48	41.16	23.20	0.6463	-4.6108
Oil	34.07	34.75	18.71	0.6379	-4.7263
Frozen CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	50.25	54.47	30.44	0.6976	-4.6260
Cartoon	42.71	41.71	23.70	0.6711	-4.8261
Pencil	42.94	41.59	23.57	0.6502	-4.6120
Oil	34.05	34.51	18.64	0.6366	-4.7259
ViT w/ (Ren et al., 2023)	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	45.49	51.01	28.04	0.6782	-4.6283
Cartoon	40.55	40.21	23.12	0.6594	-4.8264
Pencil	43.60	42.19	24.21	0.6516	-4.6103
Oil	36.31	36.19	19.53	0.6434	-4.7253
Captioning	**Trained on R+C+O**				
ViT	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	46.24	52.86	30.26	0.6836	-4.6336
Cartoon	43.58	41.88	24.14	0.6755	-4.8271
Pencil	34.18	36.55	19.77	0.6355	-4.6118
Oil	46.03	42.03	23.98	0.6752	-4.7244
Frozen CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	49.02	53.96	30.73	0.6976	-4.6282
Cartoon	44.06	42.02	24.35	0.6829	-4.8266
Pencil	35.40	36.89	20.19	0.6362	-4.6113
Oil	47.19	42.77	23.95	0.6802	-4.7231
ViT w/ (Ren et al., 2023)	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	47.46	53.09	30.42	0.6921	-4.6253
Cartoon	42.49	41.73	24.41	0.6735	-4.8279
Pencil	35.78	37.69	20.62	0.6461	-4.6020
Oil	44.86	41.51	23.86	0.6722	-4.7253
Captioning	**Trained on R+P+O**				
ViT	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	48.42	53.77	30.23	0.6879	-4.6289
Cartoon	32.80	36.12	20.76	0.6362	-4.8301
Pencil	42.19	42.17	23.70	0.6541	-4.6135
Oil	45.41	41.73	23.62	0.6741	-4.7233
Frozen CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	49.41	54.89	30.94	0.6976	-4.6251
Cartoon	33.31	37.15	20.69	0.6337	-4.8264
Pencil	44.32	43.21	24.87	0.6578	-4.6111
Oil	46.56	42.59	23.97	0.6774	-4.7241
ViT w/ (Ren et al., 2023)	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	48.84	54.29	30.55	0.6945	-4.6253
Cartoon	35.69	38.24	21.49	0.6416	-4.8253
Pencil	43.77	42.72	24.12	0.6525	-4.6103
Oil	43.27	40.33	23.46	0.6763	-4.7253
Captioning	**Trained on C+P+O**				
ViT	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	21.19	33.95	22.38	0.5799	-4.6305
Cartoon	44.05	43.38	24.47	0.6833	-4.8261
Pencil	44.10	42.21	23.94	0.6574	-4.6103
Oil	47.29	42.50	23.99	0.6779	-4.7253
Frozen CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	21.73	33.31	22.79	0.5835	-4.6252
Cartoon	44.48	43.79	24.33	0.6892	-4.8254
Pencil	44.65	42.95	24.46	0.6587	-4.6116
Oil	47.85	43.10	24.84	0.6786	-4.7248
ViT w/ (Ren et al., 2023)	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	24.51	35.06	23.63	0.6013	-4.6159
Cartoon	43.49	42.56	24.14	0.6826	-4.8265
Pencil	43.89	42.07	23.88	0.6552	-4.6119
Oil	46.51	42.03	23.78	0.6762	-4.7254

Table 5: Experimental result demonstrating the effectiveness of the domain generalization technique in image captioning tasks. Please refer to Table 18 for the results of the models trained with other setups.

VQA	Trained on R+C+P			
	Real	Cartoon	Pencil	Oil
ViT	55.65	74.11	75.29	75.88
Frozen CLIP	58.52	76.47	75.58	76.59
ViT w/ (Ren et al., 2023)	57.35	73.84	73.52	78.82
VQA	**Trained on R+C+O**			
	Real	Cartoon	Pencil	Oil
ViT	57.05	74.52	73.62	76.17
Frozen CLIP	57.94	75.58	74.53	78.23
ViT w/ (Ren et al., 2023)	57.63	74.03	76.17	74.70
VQA	**Trained on R+P+O**			
	Real	Cartoon	Pencil	Oil
ViT	55.58	71.76	76.53	76.94
Frozen CLIP	56.76	73.23	76.85	77.62
ViT w/ (Ren et al., 2023)	55.29	74.41	76.26	76.79
VQA	**Trained on C+P+O**			
	Real	Cartoon	Pencil	Oil
ViT	45.53	74.82	76.62	77.06
Frozen CLIP	47.64	75.18	77.04	77.15
ViT w/ (Ren et al., 2023)	48.82	74.85	75.94	76.56

Table 6: Experimental result demonstrating the effectiveness of domain generalization technique in VQA task. Please refer to Table 19 for the results of the models trained with other setups.

VE	Trained on R+C+P			
	Real	Cartoon	Pencil	Oil
ViT	72.20	71.30	69.11	70.04
Frozen CLIP	72.93	71.63	69.49	70.39
ViT w/ (Ren et al., 2023)	71.62	70.73	68.77	70.96
VE	**Trained on R+C+O**			
	Real	Cartoon	Pencil	Oil
ViT	71.74	71.51	68.47	69.35
Frozen CLIP	72.35	72.02	68.66	70.16
ViT w/ (Ren et al., 2023)	71.62	70.81	69.03	68.77
VE	**Trained on R+P+O**			
	Real	Cartoon	Pencil	Oil
ViT	71.76	67.30	69.22	69.61
Frozen CLIP	72.11	68.19	70.16	70.09
ViT w/ (Ren et al., 2023)	71.73	69.43	68.95	69.11
VE	**Trained on C+P+O**			
	Real	Cartoon	Pencil	Oil
ViT	58.40	70.65	69.16	69.54
Frozen CLIP	59.31	71.42	70.21	70.09
ViT w/ (Ren et al., 2023)	59.79	70.20	68.70	69.23

Table 7: Experimental result demonstrating the effectiveness of domain generalization technique in VE task. Please refer to Table 20 for the results of the models trained with other setups.

ing VOLDOGER, underscoring the need for future research in this area.

5.3 Effectiveness of Domain Generalization Techniques to Mitigate Domain Shift

Subsequently, we evaluate the effectiveness of the domain generalization techniques in mitigating the domain shift identified, as discussed in the previous section. In this experiment, we employed the domain generalization method from a previous study using a ViT encoder (Ren et al., 2023). Because this strategy focused solely on the image captioning task, we extended it to VQA and VE tasks, with the modifications detailed in Appendix B. We established two baselines for this experiment: joint training without a dedicated strategy using ViT encoders and fixed CLIP encoders.

Tables 5, 6 and 7 list the experimental results. In general, we found that using multiple source domains enhanced the out-domain performance, as indicated in red in the tables, compared to models trained on a single domain. Additionally, we discovered that the dedicated domain generalization strategy for vision-language tasks is more beneficial for out-domain performance than naive joint training. However, the implementation of such a strategy exhibited a slightly lower in-domain performance than the baselines. This highlights the potential for improvements in domain generalization techniques for vision-language tasks. We believe that the proposed VOLDOGER will play a crucial role in the development and benchmarking of this direction.

6 Conclusion

In this study, we propose a data annotation framework that leverages multimodal LLMs to construct a dataset with various styles for vision-language tasks. We created VOLDOGER, a dataset for three vision-language tasks with four different image styles by exploiting the proposed pipeline. Using VOLDOGER, we conducted extensive experiments across three tasks using various models. Our experiments confirmed the existence of a domain shift in vision-language tasks when dealing with images in different styles compared with the training data. In addition, we validated the effectiveness of the domain generalization strategy in our setup. We believe that our framework and VOLDOGER will serve as cornerstones for future research on domain generalization for vision-language tasks.

Limitations

In this section, we discuss the potential limitations of our study. First, it should be noted that the primary consideration of VOLDOGER is the stylistic domain shift of image in vision-language tasks, rather than other elements such as cultural representation of the image or linguistic differences. Nevertheless, our VOLDOGER is a dedicated dataset for evaluating and mitigating stylistic domain shift, playing a complementary role with the dataset for semantic domain shift proposed by previous study (Ren et al., 2023).

Second, the analysis presented in Appendix C regarding the distribution of labels in each dataset, as depicted in Figure 3 and 4, revealed that the distribution of the label differs from that of the original VQA and VE datasets. This is attributed to the difference between x_{ori} and x_{sty}, which is marginal in general, but can alter the label of the question or hypothesis. For instance, the example in Appendix F.3 shows a change in the label regarding the question. In particular, the answer to the question asking the position of the tennis athletes is "Yes" for x_{ori} but "No" for x_{sty}. While more meticulous verification methods for generated images, such as TIFA (Hu et al., 2023) could potentially mitigate this issue by ensuring the similarity between x_{ori} and x_{sty}, such restriction could raise the cost for overall progress. Considering this, in future work, we will focus on improving the proposed annotation method such that it considers the preservation of the label and maintains label distribution, with the consideration of the trade-off between cost and correctness.

Ethics Statement

It should be considered that LLMs and generative models may have potential biases (Gallegos et al., 2024; Zhou et al., 2024; Vice et al., 2024), leading to unintended biases in the dataset created by LLM-based annotation. Consequently, VOLDOGER may also contain several biases. These potential biases do not reflect the viewpoint of the authors. Nonetheless, we would like to note that we could not find the cases of biases in the annotated data from our manual inspection.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of CVPR*, pages 4971–4980.

Anthropic. 2024. Introducing the next generation of claude. Accessed: May 21, 2024.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *OpenAI Blog*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*, pages 440–447.

Nitay Calderon, Naveh Porat, Eyal Ben-David, Zorik Gekhman, Nadav Oved, and Roi Reichart. 2023. Measuring the robustness of natural language processing models to domain shifts. *arXiv preprint arXiv:2306.00168*.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56.

Qingchao Chen, Yang Liu, and Samuel Albanie. 2021. Mind-the-gap! unsupervised domain adaptation for text-video retrieval. In *Proceedings of AAAI*, volume 35, pages 1072–1080.

Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 521–530.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Juhwan Choi, Eunju Lee, Kyohoon Jin, and YoungBin Kim. 2024. Gpts are multilingual annotators for sequence generation tasks. In *Findings of EACL*, pages 17–40.

Ana Cláudia Akemi Matsuki de Faria, Felype de Castro Bastos, José Victor Nogueira Alves da Silva, Vitor Lopes Fabris, Valeska de Sousa Uchoa, Décio Gonçalves de Aguiar Neto, and Claudio Filipi Goncalves dos Santos. 2023. Visual question answering: A survey on techniques and common trends in recent literature. *arXiv preprint arXiv:2305.11033*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is gpt-3 a good data annotator? In *Proceedings of ACL*, pages 11173–11195.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*.

Hady Elsahar and Matthias Gallé. 2019. To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. In *Proceedings of NeurIPS*, pages 35544–35575.

Yuqi Fang, Pew-Thian Yap, Weili Lin, Hongtu Zhu, and Mingxia Liu. 2024. Source-free unsupervised domain adaptation: A survey. *Neural Networks*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of NAACL*, pages 6556–6576.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Bhanuka Manesha Samarasekara Vitharana Gamage and Lim Chern Hong. 2021. Improved ramen: towards domain generalization for visual question answering. *arXiv preprint arXiv:2109.02370*.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of NAS*.

Google. 2024a. Gemini 1.5: Our next-generation model, now available for private preview in google ai studio. Accessed: May 21, 2024.

Google. 2024b. Introducing paligemma, gemma 2, and an upgraded responsible ai toolkit. Accessed: May 21, 2024.

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 CVPR*, pages 6904–6913.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. 2006. A kernel method for the two-sample-problem.

Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, Ji Zhang, and Alberto Del Bimbo. 2021. Adavqa: Overcoming language priors with adapted margin cosine loss. In *Proceedings of IJCAI*, pages 708–714.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*, pages 770–778.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of NAACL (Industry Track)*, pages 165–190.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of ICCV*, pages 20406–20417.

Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. 2020. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of AAAI*, pages 11181–11188.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Quan Hoang Lam, Quang Duy Le, Van Kiet Nguyen, and Ngan Luu-Thuy Nguyen. 2020. Uit-viic: A dataset for the first evaluation on vietnamese image captioning. In *Proceedings of ICCCI*, pages 730–742.

Changmao Li and Jeffrey Flanigan. 2024. Task contamination: Language models may not be few-shot anymore. In *Proceedings of AAAI*, pages 18471–18480.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of ICML*, pages 19730–19742.

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 ICML*, pages 12888–12900.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023b. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of EMNLP*, pages 1487–1505.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of ACL 2004 Workshop Text Summarization Branches Out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of NeurIPS*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2024. Hello gpt-4o. Accessed: May 21, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, pages 8748–8763.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Proceedings of NeurIPS*.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of NAACL 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147.

Yuchen Ren, Zhendong Mao, Shancheng Fang, Yan Lu, Tong He, Hao Du, Yongdong Zhang, and Wanli Ouyang. 2023. Crossing the gap: Domain generalization for image captioning. In *Proceedings of CVPR*, pages 2871–2880.

Alexandra Schofield, Laure Thompson, and David Mimno. 2017. Quantifying the effects of text duplication on semantic models. In *Proceedings of EMNLP*, pages 2737–2747.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *arXiv preprint arXiv:1706.09799*.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2019. Answer them all! toward universal visual question answering models. In *Proceedings of CVPR*, pages 10472–10481.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):539–559.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Suraj Jyothi Unni, Raha Moraffah, and Huan Liu. 2023. Vqa-gen: A visual question answering benchmark for domain generalization. *arXiv preprint arXiv:2311.00807*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. 2024. Severity controlled text-to-image generative model bias manipulation. *arXiv preprint arXiv:2404.02530*.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072.

Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. 2020. Learning from extrinsic and intrinsic supervisions for domain generalization. In *Proceedings of ECCV*, pages 159–176.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. In *Findings of EMNLP*, pages 4195–4205.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP (Demo Track)*, pages 38–45.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, and Kai Lei. 2018. Multitask learning for cross-domain image captioning. *IEEE Transactions on Multimedia*, 21(4):1047–1061.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. pages 27263–27277.

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023a. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.

Mingda Zhang, Tristan Maidment, Ahmad Diab, Adriana Kovashka, and Rebecca Hwa. 2021. Domain-robust vqa with diverse datasets and methods but no target labels. In *Proceedings of CVPR*, pages 7046–7056.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023b. Llmaaa: Making large language models as active annotators. In *Findings of EMNLP*, pages 13088–13103.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of ICLR*.

Wei Zhao, Wei Xu, Min Yang, Jianbo Ye, Zhou Zhao, Yabing Feng, and Yu Qiao. 2017. Dual learning for cross-domain image captioning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 29–38.

Wentian Zhao, Xinxiao Wu, and Jiebo Luo. 2020. Cross-domain image captioning via cross-modal retrieval and model adaptation. *IEEE Transactions on Image Processing*, 30:1180–1192.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415.

Mi Zhou, Vibhanshu Abhishek, Timothy Derdenger, Jaymo Kim, and Kannan Srinivasan. 2024. Bias in generative ai. *arXiv preprint arXiv:2403.02726*.

A Evaluation on Zero-shot Performance of Multimodal Large Language Models

In this section, we present the evaluation of the zero-shot performance of recent multimodal LLMs that can perform various tasks without specific training (Yin et al., 2023). We adopted open-source models such as BLIP-2 (Li et al., 2023a), PaliGemma (Google, 2024b), and LLaVA (Liu et al., 2023), as well as proprietary models such as GPT-4 (OpenAI, 2023), Gemini (Google, 2024a), and Claude 3 (Anthropic, 2024) for a comprehensive evaluation.

The results are listed in Tables 8, 9, and 10. Overall, GPT-4o demonstrated the best performance in most cases. Additionally, we observed that several open-source models outperformed proprietary models in VQA tasks with real images but not in other tasks such as VE and image captioning. Notably, PaliGemma and LLaVA 1.5 exhibited considerably worse performance than the other models. This phenomenon may indicate the possibility of task contamination (Li and Flanigan, 2024), where these open-source models may have used VQA-v2 data during their training process[2].

The possibility of task contamination suggests that our proposed VolDoGer may not be optimal for measuring the zero-shot performance of multimodal LLMs. While we identified potential task contamination based on the performance discrepancies between PaliGemma and LLaVA 1.5 models on the VQA and VE tasks, other models, including proprietary models, may also exploit the original datasets, such as VQA-v2, SNLI-VE, and MSCOCO images, which we utilized to construct VolDoGer.

This underscores the need for a more sophisticated approach for comparing the zero-shot performance of multimodal LLMs across different styles. One potential strategy for addressing this gap is to compare the outputs produced by various models based on human preferences (Chiang et al., 2024). Specifically, this could involve crowdsourcing the collection of human preferences for different models based on specific images and ranking the models using these data.

Despite its limitations in measuring the zero-shot performance of multimodal LLMs, VolDoGer is the first dedicated dataset for domain generalization across multiple vision-language tasks with different styles. This will serve as a valuable resource for future research on domain generalization for these tasks.

[2]Note that PaliGemma clarified that they used a mixture of downstream academic datasets.

Captioning	Open-Source Models			
	Real	Cartoon	Pencil	Oil
BLIP2-FlanT5-XL	-6.395	-6.822	-6.516	-6.693
PaliGemma	-4.754	-5.868	-5.114	-5.091
LLaVA 1.5	-4.625	-4.829	-4.618	-4.725
LLaVA-NeXT w/ Vicuna-7B	-4.652	-4.883	-4.644	-4.724
LLaVA-NeXT w/ Mistral-7B	-4.698	-5.023	-4.702	-4.846
Captioning	**Proprietary Models**			
	Real	Cartoon	Pencil	Oil
GPT-4-Vision 1106-preview	-4.629	-4.827	-4.618	-4.725
GPT-4-Turbo 2024-04-09	-4.625	-4.829	-4.619	-4.725
GPT-4o 2024-05-13	-4.636	-4.836	-4.623	-4.726
Claude 3 Haiku	-4.640	-4.829	-4.624	-4.726
Claude 3 Sonnet	-4.630	-4.828	-4.617	-4.726
Claude 3 Opus	-4.639	-4.829	-4.620	-4.726
Gemini 1.0 Pro	-4.626	-4.829	-4.611	-4.725
Gemini 1.5 Flash	-4.725	-4.618	-4.827	-4.625

Table 8: Experimental result demonstrating the zero-shot performance of multimodal LLMs on image captioning task. We only report BARTScore for this experiment as matching-based metrics are less suitable for evaluating the quality of zero-shot text generation (Fu et al., 2024).

B Implementation Detail

This section presents implementation details to supplement the experimental setup described in Section 5.1. We primarily employed PyTorch (Paszke et al., 2019) and Transformers (Wolf et al., 2020) to this end. The training and inference of the fine-tuned models were performed on a single Nvidia RTX 3090 GPU, whereas the inference of multimodal large language models was conducted on a single Nvidia A100 GPU. Please refer to the source code for the annotated data and more details[3].

B.1 Image Captioning

Fine-tuned Models. For image captioning, we used ViT and CLIP encoders with a GPT-2 decoder. Specifically, we adopted *google/vit-base-patch16-224-in21k*, *openai/clip-vit-base-patch16*, and *openai-community/gpt2* from Transformers library, respectively. For BLIP, we avoided directly

[3]`https://anonymous.4open.science/r/VL_LLM_ANNO`

VQA	Open-Source Models			
	Real	Cartoon	Pencil	Oil
BLIP2-FlanT5-XL	65.29	64.41	61.18	62.92
PaliGemma	80.59	79.41	75.29	75.59
LLaVA 1.5	80.88	76.18	72.94	71.18
LLaVA-NeXT w/ Vicuna-7B	80.29	67.65	64.12	64.12
LLaVA-NeXT w/ Mistral-7B	81.76	65.88	61.18	64.41
VQA	**Proprietary Models**			
	Real	Cartoon	Pencil	Oil
GPT-4-Vision 1106-preview	75.29	67.06	59.12	62.35
GPT-4-Turbo 2024-04-09	76.47	67.65	62.94	64.71
GPT-4o 2024-05-13	77.35	82.94	79.41	78.53
Claude 3 Haiku	75.00	67.35	62.06	62.35
Claude 3 Sonnet	68.24	74.12	72.35	70.29
Claude 3 Opus	63.53	63.82	61.76	63.24
Gemini 1.0 Pro	73.23	68.24	68.23	68.82
Gemini 1.5 Flash	75.88	78.82	73.82	72.94

Table 9: Experimental result demonstrating the zero-shot performance of multimodal large language models in VQA task.

VE	Open-Source Models			
	Real	Cartoon	Pencil	Oil
BLIP2-FlanT5-XL	63.82	73.13	72.24	72.00
PaliGemma	34.33	33.91	35.02	34.79
LLaVA 1.5	33.53	29.87	33.41	32.60
LLaVA-NeXT w/ Vicuna-7B	55.76	55.25	57.95	55.18
LLaVA-NeXT w/ Mistral-7B	70.05	70.36	67.86	69.24
VE	**Proprietary Models**			
	Real	Cartoon	Pencil	Oil
GPT-4-Vision 1106-preview	65.32	70.59	70.51	71.20
GPT-4-Turbo 2024-04-09	61.75	72.43	72.58	70.05
GPT-4o 2024-05-13	71.08	73.13	72.47	70.74
Claude 3 Haiku	58.18	63.55	67.86	66.47
Claude 3 Sonnet	59.22	72.78	72.24	71.08
Claude 3 Opus	59.91	66.65	61.18	64.06
Gemini 1.0 Pro	64.63	60.32	63.13	64.29
Gemini 1.5 Flash	64.17	74.39	73.96	72.35

Table 10: Experimental result demonstrating the zero-shot performance of multimodal large language models in VE task.

applying *Salesforce/blip-image-captioning-base* as our baseline because this model had already used the MSCOCO captioning dataset for continual pre-training. Instead, we loaded the raw checkpoint of the BLIP model before pre-training. Every model was trained based on the Adam (Kingma and Ba, 2015) optimizer with a learning rate of 5e-5 for three epochs without the deployment of a scheduler. The batch size of the model was set to 16. Each input image was resized to 256×256 size and the region with 224×224 size was randomly cropped from the resized image during training. For inference, a 224×224 region was obtained from the center of the resized image. This resizing and cropping procedure was applied to each model and across all three tasks.

Domain Generalization Method. We implemented the domain generalization method we used for our experiment from scratch because there is no available source code (Ren et al., 2023). Although we followed their explanation to implement the framework, it is important to note that we used the encoded feature of the ViT encoder instead of the ResNet (He et al., 2016) model.

Zero-shot Models. We adopted *google/paligemma-3b-mix-224*, *llava-hf/llava-1.5-7b-hf*, *llava-hf/llava-v1.6-vicuna-7b-hf*, and *llava-hf/llava-v1.6-mistral-7b-hf* from Transformers as PaliGemma, LLaVA-1.5, LLaVA-NeXT w/ Vicuna, and LLaVA-NeXT w/ Mistral, respectively, in our experiments. We used slightly different input prompts for the open-source and proprietary models because proprietary models offer system prompts. For open-source models, we used a relatively simple prompt, which is "Provide a detailed description of the given image in one sentence." For proprietary models involving GPT-4, Claude, and Gemini, we applied the following system prompt: "You are a helpful AI assistant that helps people generate captions for their images. Your output should be a single sentence that describes the image. Do not generate any inappropriate or accompanying text." The input prompt was set to "Please generate a caption for this image. Please generate the result in the form of Caption: <your caption here>".

Evaluation Metric. The NLG-EVAL library (Sharma et al., 2017) was used to measure the BLEU, ROUGE, and METEOR metrics. We reported the average of BLEU-1, 2, 3, and 4 scores as BLEU score. For BERTScore and BARTScore, we adopted the *bert-base-uncased* and *facebook/bart-large-cnn* models, respectively.

B.2 Visual Question Answering

Fine-tuned Models. We used identical models for ViT and CLIP image encoders. For the BERT text encoder, we adopted the *bert-base-uncased* model. Each output feature produced by the image and text encoders with a vector size of 768 was concatenated into a single feature with a size of 1536, and was fed into the classifier with a single ReLU activation. For the BLIP model, we used the raw checkpoint instead of the *Salesforce/blip-vqa-base*. We trained each model with a learning rate of 5e-5 for 10 epochs using the Adam optimizer, with early stopping based on the accuracy of the validation set.

Domain Generalization Method. We used the image caption offered by the MSCOCO captioning dataset because VQA-v2 dataset was also built on images from MSCOCO.

Zero-shot Models. For open-source models, we used the following simple prompt: "Question: based on the image, {question}? Answer with yes or no." For proprietary models, we applied the following system prompt: "You are a helpful AI assistant that helps visual question answering tasks.", while the input prompt was set to "Please answer the question below based on the given image. Start the response with Yes or No. Question: {question}?"

B.3 Visual Entailment

Fine-tuned Models. We used a model structure identical to that used for VQA task. The models were trained using the Adam optimizer for three epochs with a learning rate of 5e-5.

Domain Generalization Method. We used the text premise offered by SNLI-VE dataset as a description of a given image, as they are the captions from Flickr30k dataset, the source of the image of SNLI-VE.

Zero-shot Models. We used the following simple prompt for open-source models: "Statement: {hypothesis} Determine if the statement is true, false, or undetermined based on the image. Answer with true, false, or undetermined." For proprietary models, we applied the following system prompt: "You are a helpful AI assistant that helps visual entailment tasks.", and the input prompt applied was "Does the given hypothesis entail the image? Start the response with True, False, or Undetermined. Hypothesis: {hypothesis}"

B.4 Data Annotation

For data annotation, we used the GPT-4o (OpenAI, 2024) model as our $\mathcal{M}$. The model version was *GPT-4o-2024-05-13*. We set every parameter, including the top-p and temperature as default. We set the patience of error to 10, and the data that exceeded this patience were omitted from the annotation procedure. Prompts for the annotation process such as P_{ID} are provided in Appendix G. In addition, we used DALL-E 3 (Betker et al., 2023) as the image generation model $\mathcal{G}$. Note that other image generation models such as Stable Diffusion (Esser et al., 2024) can also be used as $\mathcal{G}$ instead of DALL-E 3. The overall data annotation procedure costs approximately USD 1,800.

C Dataset Specification

In this section, we provide more detailed information on the VOLDOGER. Additionally, Figure 5 suggests the result t-SNE visualization (Van der Maaten and Hinton, 2008) for each domain of three tasks, especially demonstrating visual domain gaps.

C.1 VOLDOGER-CAP

Table 11 lists the number of images for each style in VOLDOGER-CAP. Each style contains approximately 3,850 images, with five different captions for each image.

Captioning	Train	Validation	Test	Total
Real	2695	924	231	3850
Cartoon	2695	924	231	3850
Pencil	2694	923	231	3848
Oil	2694	924	231	3849

Table 11: The amount of images for each style in VOLDOGER-CAP.

C.2 VOLDOGER-VQA

Tables 12 and 13 present the number of images and questions as well as the domain gap for each style in VOLDOGER-VQA. Figure 3 presents the number of labels for each split.

C.3 VOLDOGER-VE

Tables 14 and 15 present the number of images, hypotheses, and the domain gap for each style in

VQA Images	Train	Valid	Test	Total
Real	2091	711	182	2984
Cartoon	2090	710	182	2982
Pencil	2090	711	182	2983
Oil	2091	711	182	2984
VQA Questions	Train	Valid	Test	Total
Real	4120	1452	340	5912
Cartoon	4118	1451	340	5909
Pencil	4118	1452	340	5910
Oil	4120	1452	340	5912

Table 12: The number of images and questions for each style in VOLDOGER-VQA.

	R	**C**	**P**	**O**	
R	-	0.0024	0.0026	0.0026	
C	0.0127	-	0.0016	0.0016	**Average**
P	0.0165	0.0109	-	0.0014	0.0020
O	0.0124	0.0091	0.0106	-	0.0120

Table 13: Domain gap of each style in VOLDOGER-VQA, measured with MMD by ResNet and BERT output vectors. Orange figures denote the visual domain gap, and blue figures represent the linguistic domain gap.

VE Images	Train	Valid	Test	Total
Real	619	77	78	774
Cartoon	618	77	78	773
Pencil	619	77	78	774
Oil	619	77	78	774
VE Hypotheses	Train	Valid	Test	Total
Real	7673	967	868	9508
Cartoon	7670	966	867	9503
Pencil	7665	967	868	9500
Oil	7666	967	868	9501

Table 14: Number of images and questions for each style in VOLDOGER-VE.

	R	**C**	**P**	**O**	
R	-	0.0060	0.0067	0.0062	
C	0.0109	-	0.0042	0.0044	**Average**
P	0.0146	0.0109	-	0.0038	0.0052
O	0.0106	0.0087	0.0104	-	0.0110

Table 15: Domain gap of each style in VOLDOGER-VE, measured with MMD by ResNet and BERT output vectors. Orange figures denote the visual domain gap, and blue figures represent the linguistic domain gap.

VOLDOGER-VE. Figure 4 presents the number of labels for each split.

D Ablation Study

In this section, we conduct an ablation study that validates the effectiveness of label verification and re-annotation in VQA and VE tasks.

D.1 Manual Analysis on Label Verification

First, we manually investigated the results of label verification and label re-annotation. We selected the test split of three styles in the VQA task as representatives. Subsequently, we gathered data with labels that differed from those in the real photo domain. As a result, we acquired 127 questions from the cartoon drawing domain, 134 questions from the pencil drawing domain, and 130 questions from the oil painting domain. We then examined the annotation results to determine their acceptability. We found that 26 questions from the cartoon drawing domain, 24 questions from the pencil drawing domain, and 25 questions from the oil painting domain were unacceptable and falsely annotated, accounting for less than 20% of each domain.

Furthermore, we observed several tendencies in LLM-based annotations. For instance, the LLM predominantly suggested "No" for subjective questions such as "Is the weather cold?", "Is this man happy?", or "Is the boy good at this game?". Moreover, the LLM struggled with questions asking about the professionalism of a game, such as "Is this a major league game?". We aim to investigate these tendencies more thoroughly in future work. Additionally, This analysis is included in the dataset repository as a report, providing a broad perspective and assisting future studies.

D.2 Experiment on Answer Verification

Second, we conducted an ablation experiment by directly assigning labels from the real photo domain, thereby excluding the answer verification process. We created an ablation training set based on this setup and trained three VQA models for each style, evaluating their performance on in-domain test sets. The results are presented in Table 16. The findings suggest that directly assigning

VQA	Cartoon	Pencil	Oil
w/ Answer Verification (Ours)	75.23	75.29	77.35
w/o Answer Verification (Ablation)	71.17	73.23	75.58

Table 16: The result of ablation experiment that excludes answer verification process from our framework.

labels from the real photo domain to other domains can harm model performance, as the distinction between real and generated images, along with their labels, acts as noisy labels.

In conclusion, both the manual analysis and the experimental results support the significance of the answer and label verification and re-annotation procedure we proposed in Section 3.3 and 3.4.

E Further Experimental Result

E.1 Domain Shift of Model in Image Captioning

Captioning	**Trained on Cartoon Drawing**				
ViT	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	21.64	35.07	21.82	0.5916	-4.6258
Cartoon	42.53	41.86	23.38	0.6721	-4.8267
Pencil	31.50	35.63	18.79	0.6267	-4.6112
Oil	30.66	33.39	17.32	0.6270	-4.7253
CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	20.20	32.90	19.23	0.5858	-4.6253
Cartoon	38.66	39.99	21.72	0.6595	-4.8271
Pencil	24.04	30.14	15.93	0.6036	-4.6126
Oil	27.69	30.97	15.77	0.6105	-4.7255
BLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	21.21	32.60	22.96	0.5866	-4.6698
Cartoon	41.75	40.26	25.28	0.6822	-4.8737
Pencil	33.73	34.49	20.84	0.6313	-4.6439
Oil	34.83	34.71	18.92	0.6380	-4.7294
Captioning	**Trained on Pencil Drawing**				
ViT	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	21.61	34.18	22.31	0.5933	-4.6253
Cartoon	35.50	38.05	20.66	0.6403	-4.8264
Pencil	42.87	41.52	23.18	0.6481	-4.6106
Oil	33.92	34.56	18.38	0.6475	-4.7253
CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	20.73	32.72	18.59	0.5711	-4.6253
Cartoon	30.28	34.02	17.64	0.6104	-4.8264
Pencil	39.88	39.42	21.37	0.6298	-4.6103
Oil	30.67	32.22	16.52	0.6261	-4.7253
BLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	19.18	29.52	22.64	0.5735	-4.6752
Cartoon	34.30	34.04	21.47	0.6479	-4.8780
Pencil	42.14	38.74	23.93	0.6537	-4.6415
Oil	33.97	33.41	19.17	0.6406	-4.7284
Captioning	**Trained on Oil Painting**				
ViT	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	19.20	29.74	21.41	0.5684	-4.6254
Cartoon	33.76	35.58	21.37	0.6350	-4.8274
Pencil	34.34	34.85	20.60	0.6361	-4.6111
Oil	46.97	42.39	23.75	0.6759	-4.7253
CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	19.33	29.83	20.30	0.5705	-4.6251
Cartoon	32.13	34.14	19.56	0.6237	-4.8262
Pencil	31.34	32.51	18.35	0.6268	-4.6103
Oil	46.11	42.09	23.05	0.6693	-4.7253
BLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	21.12	30.38	22.26	0.5818	-4.6253
Cartoon	34.41	36.09	21.20	0.6335	-4.8264
Pencil	35.30	35.32	20.55	0.6373	-4.6105
Oil	46.67	41.18	25.01	0.6833	-4.7306

Table 17: Supplementary experimental result demonstrating domain shift on image captioning task.

E.2 Effectiveness of Domain Generalization in Image Captioning Task

Captioning	Trained on R+C						Trained on R+P				
ViT	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	44.82	51.13	28.46	0.6838	-4.6268	Real	44.74	51,14	28.30	0.6828	-4.6261
Cartoon	41.95	41.64	23.35	0.6701	-4.8279	Cartoon	36.19	38.13	20.56	0.6410	-4.8279
Pencil	31.48	36.07	18.76	0.6275	-4.6130	Pencil	43.01	41.91	23.28	0.6519	-4.6116
Oil	30.09	32.99	17.23	0.6281	-4.7254	Oil	31.08	33.22	17.68	0.6287	-4.7277
Frozen CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	50.10	54.53	30.61	0.6977	-4.6260	Real	49.20	54.48	30.54	0.6969	-4.6253
Cartoon	42.06	41.87	23.35	0.6718	-4.8264	Cartoon	35.69	37.50	20.61	0.6353	-4.8264
Pencil	30.64	35.48	18.75	0.6288	-4.6116	Pencil	43.51	41.17	23.02	0.6510	-4.6105
Oil	26.81	31.54	16.40	0.6170	-4.7253	Oil	29.12	32.14	17.66	0.6222	-4.7253
ViT w/ (Ren et al., 2023)	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	39.97	48.21	27.84	0.6769	-4.6283	Real	46.96	35.16	29.50	0.6905	-4.6284
Cartoon	42.38	42.04	23.04	0.6649	-4.8264	Cartoon	36.65	38.56	21.61	0.6406	-4.8236
Pencil	31.72	36.38	19.17	0.6318	-4.6103	Pencil	42.05	40.79	23.23	0.6517	-4.6124
Oil	30.05	32.73	17.52	0.6306	-4.7235	Oil	31.59	34.17	17.70	0.6414	-4.7235
Captioning	**Trained on R+O**						**Trained on C+P**				
ViT	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	45.49	52.24	28.98	0.6947	-4.6274	Real	20.65	33.31	22.14	0.5762	-4.6293
Cartoon	32.04	34.64	19.10	0.6265	-4.8269	Cartoon	42.54	42.18	23.24	0.6616	-4.8267
Pencil	32.73	34.08	18.66	0.6272	-4.6134	Pencil	43.12	41.14	23.62	0.6469	-4.6105
Oil	45.81	42.29	23.34	0.6749	-4.7254	Oil	34.14	34.79	18.97	0.6394	-4.7254
Frozen CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	48.64	53.52	30.03	0.6938	-4.6292	Real	20.99	34.66	22.45	0.5771	-4.6290
Cartoon	33.35	34.42	19.86	0.6297	-4.8266	Cartoon	42.82	42.20	23.64	0.6733	-4.8264
Pencil	31.74	34.53	18.88	0.6323	-4.6122	Pencil	43.09	41.64	23.72	0.6519	-4.6103
Oil	45.68	42.04	23.05	0.6711	-4.7353	Oil	34.50	34.91	19.12	0.6353	-4.7255
ViT w/ (Ren et al., 2023)	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	44.03	50.75	28.22	0.6811	-4.6259	Real	23.89	35.50	22.89	0.6353	-4.6253
Cartoon	34.42	35.32	20.75	0.6366	-4.8264	Cartoon	41.71	40.62	23.12	0.6445	-4.8265
Pencil	34.99	34.85	19.92	0.6324	-4.6112	Pencil	42.51	41.55	23.37	0.6483	-4.6111
Oil	44.70	41.43	22.65	0.6721	-4.7224	Oil	35.88	35.42	19.20	0.6409	-4.7253
Captioning	**Trained on C+O**						**Trained on R+P**				
ViT	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	20.66	32.13	22.10	0.5763	-4.6265	Real	20.24	32.22	21.67	0.5778	-4.6256
Cartoon	43.20	42.48	24.13	0.6781	-4.8266	Cartoon	32.24	34.71	20.13	0.6285	-4.8273
Pencil	35.17	36.11	20.14	0.6316	-4.6109	Pencil	44.23	42.28	24.02	0.6547	-4.6106
Oil	46.74	42.35	23.69	0.6789	-4.7264	Oil	47.17	43.04	23.81	0.6761	-4.7253
Frozen CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	20.62	32.51	22.62	0.5815	-4.6257	Real	20.31	31.95	21.08	0.5782	-4.6267
Cartoon	43.69	42.85	23.83	0.6816	-4.8264	Cartoon	34.49	37.22	21.08	0.6380	-4.8268
Pencil	35.60	36.19	20.49	0.6345	-4.6159	Pencil	44.42	42.02	23.95	0.6549	-4.6128
Oil	46.06	42.08	24.22	0.6801	-4.7245	Oil	46.94	43.37	24.13	0.6788	-4.7256
ViT w/ (Ren et al., 2023)	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
Real	22.57	32.86	22.76	0.5829	-4.6253	Real	21.77	32.98	22.40	0.5839	-4.6253
Cartoon	42.76	42.04	23.81	0.6751	-4.8268	Cartoon	36.58	38.23	21.93	0.6402	-4.8266
Pencil	36.83	36.45	20.75	0.6429	-4.6103	Pencil	43.04	41.32	23.39	0.6498	-4.6133
Oil	46.88	42.77	23.58	0.6766	-4.7253	Oil	47.01	42.80	23.67	0.6737	-4.7261

Table 18: Supplementary experimental result demonstrating the effectiveness of domain generalization technique on image captioning task. This table presents the result of the model trained with two source domains, instead of that of Table 5 that leveraged three source domains.

E.3 Experimental Results in Visual Question Answering Task

VQA	**Trained on Cartoon Drawing**				**Trained on Pencil Drawing**				**Trained on Oil Painting**			
	Real	Cartoon	Pencil	Oil	Real	Cartoon	Pencil	Oil	Real	Cartoon	Pencil	Oil
ViT	42.39	75.23	67.88	68.04	41.79	68.42	75.29	65.29	43.82	61.56	64.70	77.35
CLIP	44.72	76.47	69.21	67.64	43.23	68.19	75.88	66.17	44.41	62.33	65.84	78.82
BLIP	45.16	78.52	68.92	69.48	43.58	69.54	77.64	67.53	44.70	63.41	67.56	79.71
VQA	**Trained on R+C**				**Trained on R+P**				**Trained on R+O**			
	Real	Cartoon	Pencil	Oil	Real	Cartoon	Pencil	Oil	Real	Cartoon	Pencil	Oil
ViT	56.88	74.50	73.88	75.15	54.62	72.79	76.82	75.91	53.52	72.53	72.24	76.35
Frozen CLIP	54.68	74.76	72.18	75.06	55.68	73.62	77.35	75.88	55.10	72.84	70.21	76.93
ViT w/ (Ren et al., 2023)	54.59	74.29	74.53	75.47	53.84	74.24	76.24	76.79	52.82	73.79	75.47	76.03
VQA	**Trained on C+P**				**Trained on C+O**				**Trained on P+O**			
	Real	Cartoon	Pencil	Oil	Real	Cartoon	Pencil	Oil	Real	Cartoon	Pencil	Oil
ViT	44.53	74.44	75.76	76.42	45.12	74.82	74.11	76.44	44.85	74.01	76.68	76.53
Frozen CLIP	45.88	74.88	76.21	76.56	45.76	75.15	74.32	76.38	45.12	74.15	77.05	77.03
ViT w/ (Ren et al., 2023)	46.47	74.59	74.93	76.94	45.98	74.53	75.29	77.06	45.29	74.88	76.53	76.47

Table 19: Supplementary experimental result demonstrating the domain shift and effectiveness of domain generalization technique on VQA task.

E.4 Experimental Results in Visual Entailment Task

VE	**Trained on Cartoon Drawing**				**Trained on Pencil Drawing**				**Trained on Oil Painting**			
	Real	Cartoon	Pencil	Oil	Real	Cartoon	Pencil	Oil	Real	Cartoon	Pencil	Oil
ViT	55.14	68.95	64.01	63.85	55.32	62.59	69.18	65.16	55.79	63.43	65.20	71.47
CLIP	54.56	69.81	65.08	64.49	56.24	63.84	69.70	64.12	55.23	62.94	64.93	71.87
BLIP	49.88	63.18	58.29	57.21	48.15	61.13	65.09	61.04	46.82	60.21	61.96	71.89
VE	**Trained on R+C**				**Trained on R+P**				**Trained on R+O**			
	Real	Cartoon	Pencil	Oil	Real	Cartoon	Pencil	Oil	Real	Cartoon	Pencil	Oil
ViT	72.83	69.89	65.78	65.52	69.70	64.28	68.89	65.38	72.02	65.32	64.71	67.96
Frozen CLIP	73.27	70.01	65.24	66.47	73.38	66.83	69.10	65.23	72.11	64.21	65.47	69.70
ViT w/ (Ren et al., 2023)	72.51	69.71	66.11	67.72	72.60	65.88	68.74	67.47	72.17	65.79	66.51	68.77
VE	**Trained on C+P**				**Trained on C+O**				**Trained on P+O**			
	Real	Cartoon	Pencil	Oil	Real	Cartoon	Pencil	Oil	Real	Cartoon	Pencil	Oil
ViT	55.28	70.55	68.66	67.96	55.97	70.63	68.08	68.43	55.06	66.32	68.69	68.97
Frozen CLIP	56.22	70.73	69.20	66.89	56.28	69.20	67.89	70.62	55.49	67.55	69.12	70.37
ViT w/ (Ren et al., 2023)	56.45	70.82	68.12	68.57	56.49	70.12	68.31	69.44	55.99	68.04	68.61	68.85

Table 20: Supplementary experimental result demonstrating the domain shift and effectiveness of domain generalization technique on VE task.

F Additional Examples of Annotated Data

F.1 Additional Examples on Image Prompt

In this section, we present an original image x_{ori} with decomposed p_{ori} and its transformed versions x_{sty} and p_{sty} generated through the annotation process described in Section 3. The expressions for p_{sty} that contribute to the generation of the stylized image are boldfaced.

F.1.1 Cartoon Drawing Style Images with Prompts

Original Image with p_{ori}	Stylized Image with p_{sty}
 Create an image of a youth soccer game in a vibrant, well-lit outdoor setting. Focus on a young boy, wearing a maroon jersey, black shorts, and black socks with white stripes, as he kicks a yellow and black soccer ball. His hair is light and short, and he appears to be mid-action, showcasing a moment of dynamic movement. Surrounding him are other young children also playing soccer, dressed in various jerseys, primarily orange and white. The background features multiple small, arched soccer goals with a netting structure, hinting at a friendly, organized game environment. Lush green grass covers the field, and in the distance, trees with autumn foliage provide a scenic backdrop. The overall atmosphere should convey the excitement and energy of a youth soccer match on a sunny day.	 Create a **cartoon-style** image of a youth soccer game in a vibrant, well-lit outdoor setting. Focus on a young boy, wearing a maroon jersey, black shorts, and black socks with white stripes, as he kicks a yellow and black soccer ball. His light, short hair and dynamic movement should be depicted with **exaggerated, playful features typical of cartoons**. Surrounding him are other young children also playing soccer, dressed in various jerseys, primarily orange and white. The background features multiple small, arched soccer goals with a netting structure, emphasizing a friendly, organized game environment. The field is illustrated with lush green grass, and in the distance, **colorful** autumn trees provide a scenic backdrop. The overall atmosphere should capture the excitement and energy of a youth soccer match on a sunny day, **with vibrant colors and whimsical details**.
 Create an image of a baseball player captured in mid-action swinging a bat. The player is wearing a white uniform with blue accents, notably with the number 51 and the name "ICHIRO" on the back. He is also wearing a black helmet, black socks, and black cleats. The scene takes place on a baseball field with a dirt basepath, green grass, and partially visible chalk lines. The player's stance and movement indicate a powerful swing, and his body is slightly bent forward with one leg stepping into the swing. The background should include the baseball field's elements subtly blurred to maintain the focus on the player. The overall atmosphere should convey the intensity and dynamism of the moment.	 Create a **cartoon-style** image of a baseball player captured in mid-action swinging a bat. The player is depicted wearing a white uniform with blue accents, with the number 51 and the name "ICHIRO" on the back. He is also wearing a black helmet, black socks, and black cleats. The scene takes place on a cartoon baseball field with a dirt basepath, green grass, and **animated** chalk lines. The player's stance and movement indicate a powerful swing, and his body is slightly bent forward with one leg stepping into the swing. The background includes **stylized** elements of a baseball field, rendered with **exaggerated features and vibrant colors**, subtly blurred to maintain the focus on the player. The overall atmosphere should convey the intensity and dynamic action in a **whimsical, cartoonish manner**.
 Create an image of a baseball player posing on a professional baseball field. The player is wearing a white baseball jersey with "CANADA" written across the chest in red letters, and a matching cap with a red maple leaf emblem. The player is pointing toward the camera with a baseball in hand, and their other hand is holding a glove. The background consists of a well-maintained baseball field, complete with bases, a pitcher's mound, and surrounding stadium seating filled with spectators. The weather is clear with scattered clouds, and lush green trees can be seen beyond the outfield. The atmosphere should be vibrant and playful, capturing the excitement of a baseball game day.	 Create a **cartoon drawing style** image of a baseball player posing on a professional baseball field. The player is wearing a white baseball jersey with "CANADA" written across the chest in red letters, and a matching cap with a red maple leaf emblem. The player is pointing toward the camera with a baseball in hand, and their other hand is holding a glove. The background consists of a well-maintained baseball field, complete with bases, a pitcher's mound, and surrounding stadium seating filled with spectators. The weather is clear with scattered clouds, and lush green trees can be seen beyond the outfield. The atmosphere should be vibrant and playful, capturing the excitement of a baseball game day with **cartoonish, exaggerated features and vivid colors**.

F.1.2 Pencil Drawing Style Images with Prompts 1205

Original Image with p_{ori}	Stylized Image with p_{sty}
 Create an image of an elephant in an outdoor setting, captured in a photorealistic style. The elephant is standing on a dirt ground with grass and vegetation in the background. It is interacting with a large ball, using its trunk to balance and play with the ball. The scene is set in a zoo or a wildlife sanctuary, with a sturdy fence in the background that includes metal posts and wires. Large rocks are present around the enclosure, enhancing the natural and rugged feel. The lighting suggests a sunny day, illuminating the elephant and casting shadows on the ground. The overall atmosphere should convey a sense of playful interaction and natural habitat for the elephant.	 Create a **pencil drawing** of an elephant in an outdoor setting, captured with **fine details and shading**. The elephant is standing on a dirt ground with grass and vegetation in the background. It is interacting with a large ball, using its trunk to balance and play with it. The scene is set in a zoo or a wildlife sanctuary, depicted with a sturdy fence in the background that includes metal posts and wires. Large rocks are present around the enclosure, enhancing the natural and rugged feel. **The drawing should incorporate detailed cross-hatching and intricate line work to create depth and texture**. The overall atmosphere should convey a sense of playful interaction and natural habitat for the elephant.
 Create an image of a baseball game scene with a focus on a player standing near the edge of the field. The player is wearing a dark jersey, white pants, and a helmet, holding a bat under their arm. They are positioned on the dirt track looking toward the field, with a helmet resting on the ground nearby. The background features a baseball field with green grass, and several people are standing on the field, wearing casual attire consisting of T-shirts and shorts. There is a trash can and sports equipment scattered on the field. The stands in the background are filled with spectators. A netting separates the viewer from the action on the field, adding to the environment of a live baseball game. The atmosphere should capture the excitement and casual moments of a baseball game.	 Create a **pencil drawing** of a baseball game scene with a focus on a player standing near the edge of the field. The player is wearing a dark jersey, white pants, and a helmet, holding a bat under their arm. They are positioned on the dirt track looking toward the field, with a helmet resting on the ground nearby. The background features a baseball field with green grass, and several people are standing on the field, wearing casual attire consisting of T-shirts and shorts. There is a trash can and sports equipment scattered on the field. The stands in the background are filled with spectators. A netting separates the viewer from the action on the field, adding to the environment of a live baseball game. **The entire scene should be rendered with detailed lines, cross-hatching, and shading to emphasize the pencil drawing style**, capturing the excitement and casual moments of a baseball game.
 Create an image of a young boy playing baseball outdoors on a sunny day. The scene is set in a grassy field with tall buildings in the background. The boy is wearing a red baseball cap, a white t-shirt, and blue shorts. He is in mid-swing, holding a bat as if he's about to hit a baseball. The sky is clear with a few scattered clouds, and several cars are parked near the buildings in the background. The overall atmosphere should capture a warm, sunny day with the boy engrossed in his game amidst an urban setting. The image should reflect a casual, recreational vibe, emphasizing the contrast between the natural, grassy field and the tall, urban buildings.	 Create a **pencil drawing** of a young boy playing baseball outdoors on a sunny day. The scene is set in a grassy field with tall buildings in the background. The boy is wearing a red baseball cap, a white t-shirt, and blue shorts. He is in mid-swing, holding a bat as if he's about to hit a baseball. The sky is clear with a few scattered clouds, and several cars are parked near the buildings in the background. The overall atmosphere should capture a warm, sunny day with the boy engrossed in his game amidst an urban setting. **The pencil drawing should include detailed line work, shading, and cross-hatching to give depth and texture**, capturing the contrast between the natural, grassy field and the tall, urban buildings.

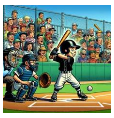
F.1.3 Oil Painting Style Images with Prompts

Original Image with p_{ori}	Stylized Image with p_{sty}
 Create an image set in an outdoor park under the shade of tall, green trees. In the foreground, capture two men standing close to each other. The man on the left is wearing a yellow t-shirt with bold text across the chest and light-colored shorts. The man on the right is wearing a white shirt and shorts, along with a white cap, and has his arm around the other man's shoulder. Both are smiling or displaying a friendly demeanor. In the background, depict a third man wearing a white shirt and black shorts, also with a cap, who is walking away while looking down at something in his hands, such as a notebook or a phone. The ground is a mix of dirt and grass, and the sunlight filters through the tree branches, creating dappled shadows. Additionally, there should be some buildings or structures faintly visible further back, adding depth to the scene. The overall atmosphere should be casual and friendly, suggesting a leisurely day in the park.	 Create an **oil painting** set in an outdoor park under the shade of tall, green trees. In the foreground, depict two men standing close to each other with warm, textured brushstrokes. The man on the left is wearing a yellow t-shirt with bold text across the chest and light-colored shorts, and the man on the right is wearing a white shirt, shorts, and a white cap, with his arm around the other man's shoulder. Both are portrayed with friendly expressions, **rendered with rich, vivid colors typical of oil paintings**. In the background, include a third man in a white shirt and black shorts, also wearing a cap, walking away while looking down at something in his hands, such as a notebook or a phone. The ground should be **painted** with a mix of earthy tones depicting dirt and grass, with sunlight filtering through the tree branches, creating dappled shadows. Additionally, some buildings or structures should be faintly visible further back, adding depth to the scene. The overall atmosphere should convey a casual and friendly leisurely day in the park, **with the warmth and depth characteristic of an oil painting**.
 Create an image of a tennis player in the middle of executing a powerful serve during a match. The player is wearing a bright orange outfit and white shoes, with a red headband. The player is positioned on a blue and green tennis court, with one arm extended upward, holding the racquet ready to strike the ball. The stance and motion should convey intensity and athleticism. In the background, there is a stadium filled with spectators watching the match, with some sections covered by tarps. The scene should capture the dynamic energy and focus of a professional tennis match in a large, well-lit arena.	 Create an **oil painting** of a tennis player in the middle of executing a powerful serve during a match. The player is wearing a bright orange outfit and white shoes, with a red headband, **all depicted with the textured brushstrokes and rich colors characteristic of oil painting.** The player is positioned on a **vibrant** blue and green tennis court, with one arm extended upward, holding the racquet ready to strike the ball. The stance and motion should convey intensity and athleticism, **captured with dynamic brushwork**. In the background, a stadium filled with spectators is **illustrated with a blend of detailed and impressionistic techniques**, showcasing their engagement and anticipation. Some sections of the stands are covered by tarps. The scene should evoke the dynamic energy and focus of a professional tennis match in a large, well-lit arena, **with an emphasis on the vivid, expressive style of an oil painting**.
 Create an image of a dynamic indoor handball match in progress. In the foreground, a player in a bright green jersey and white shorts is captured in mid-air as he attempts a powerful shot at the goal. He holds the ball in his right hand, showcasing his athleticism. To his left, two players dressed in red jerseys and white shorts are intensely focused on the play, one of them actively engaged in defense. In the right foreground, a referee in an orange shirt and black pants, with the number 16 on his back, is standing with his whistle ready to ensure fair play. The crowd in the background is seated in a dimly lit arena, watching the action with keen interest. Prominent banners and advertisements, including one with the text "VAL de MARNE Conseil général" and another for "lemarrane.com," are displayed along the sides of the court, enhancing the realistic atmosphere of a professional handball game. The flooring is a polished wooden surface, capturing the energy and intensity of the match.	 Create an **oil painting** of a dynamic indoor handball match in progress. In the foreground, a player in a bright green jersey and white shorts is depicted in mid-air, attempting a powerful shot at the goal with the ball in his right hand. The **painting** should capture his athleticism and motion with **expressive brushstrokes**. To his left, two players in red jerseys and white shorts are intensely focused on the play, one of them actively engaged in defense. On the right, a referee in an orange shirt and black pants, with the number 16 on his back, stands with his whistle ready to ensure fair play. The crowd in the background is seated in a dimly lit arena, watching the action with keen interest, **rendered with artistic details**. Prominent banners and advertisements, including one with the text "VAL de MARNE Conseil général" and another for "lemarrane.com," are **painted** along the sides of the court, enhancing the realistic atmosphere of a professional handball game. The polished wooden flooring should be depicted with **rich textures**, capturing the energy and intensity of the match **through the depth and warmth typical of an oil painting**.

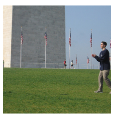
F.2 Additional Examples on Image Captioning Task

Original Data	Annotated Data
 • A group of basketball players on court during a game • Basketball players in the process of making and defending a basket during a basketball game in an arena. • A group of basketball players in the court as crowd looks • Some men playing basketball with some fans watching • A group of men playing basketball against each other.	 • A lively cartoon scene of basketball players on the court during an intense game with a packed arena. • Animated basketball players in mid-action, defending and attempting a shot in a vibrant, crowded indoor arena. • Cartoon-style basketball players energetically competing on the court as a colorful crowd watches. • Dynamic image of men playing basketball in an animated style, with enthusiastic fans cheering in the background. • Animated depiction of a group of men engaged in a basketball game, surrounded by a lively audience in a large arena.
 • A group of girls on a field playing soccer. • A group of women playing soccer on field with people watching. • Two women chasing after a soccer ball on a field. • Two girls on opposite teams competing for the soccer ball. • Two teams playing soccer while people are watching.	 • Two female soccer players in dynamic motion as they compete for the ball on a crowded field. • An intense women's soccer match, skillfully illustrated in pencil, with spectators cheering in the background. • Two determined athletes from opposing teams vying for control of the ball during a fierce soccer game. • A high-energy soccer match with two women battling for possession, surrounded by an enthusiastic crowd. • A competitive soccer scene, showing two women in action and an audience engrossed in the game, all rendered in intricate pencil detail.
 • A woman swings her tennis racket at a tennis ball. • A lady wearing white shoes and in a black outfit is playing tennis. • A woman extends her arm to hit a tennis ball. • A beautiful young woman hitting a tennis ball with a racquet. • A woman in a green tennis dress and white sneakers playing tennis on a court.	 • A woman in a green and white tennis dress swings her racket at a tennis ball, captured in a vibrant oil painting style. • An athlete, wearing white sneakers and a dark green outfit, is painted mid-action while playing tennis. • A depiction of a woman extending her arm to strike a tennis ball with dynamic brush-work. • A beautiful young woman hits a tennis ball with a racket in an oil-painted scene. • On an outdoor court, a woman in a green tennis dress and white sneakers engages in a tennis match, rendered with lush, textured strokes.

F.3 Additional Examples on Visual Question Answering Task

Original Data	Annotated Data	Original Data	Annotated Data
 • Question: Did he hit that ball? • Answer: No	 • Question: Did he strike the ball? • Answer: No	 • Question: Did a lot of people show up for the game? • Answer: No	 • Question: Was there a large crowd at the game? • Answer: No
 • Question: Does the boy have his head stuck in the net? • Answer: No	 • Question: Is the boy's head caught in the net? • Answer: No	 • Question: Is there a disabled person? • Answer: Yes	 • Question: Is there a person with a disability? • Answer: Yes
 • Question: Are the guys in blue wearing two different socks? • Answer: Yes	 • Question: Do the men in blue have mismatched socks? • Answer: No	 • Question: Is the girls right arm in an awkward position? • Answer: Yes	 • Question: Is the girl's right arm positioned awkwardly? • Answer: No

F.4 Additional Examples on Visual Entailment Task

Original Data	Annotated Data	Original Data	Annotated Data
 • Hypothesis: Adults are playing frisbee • Label: Contradiction	 • Hypothesis: Grown-ups are tossing a frisbee around. • Label: Contradiction	 • Hypothesis: Two sports players are sprinting towards the ball. • Label: Entailment	 • Hypothesis: Two athletes are racing toward the ball. • Label: Entailment
 • Hypothesis: The football coach did not see the play being made. • Label: Contradiction	 • Hypothesis: The football coach missed seeing the play happen. • Label: Contradiction	 • Hypothesis: A player injured and getting treated by the team doctor. • Label: Neutral	 • Hypothesis: An athlete is being attended to by the team doctor for an injury. • Label: Neutral
 • Hypothesis: A little boy wearing blue takes a kick at a soccer ball. • Label: Entailment	 • Hypothesis: A young boy dressed in blue attempts to kick a soccer ball. • Label: Entailment	 • Hypothesis: The man is in Pakistan kicking a soccer ball. • Label: Contradiction	 • Hypothesis: The person is playing soccer in Pakistan. • Label: Contradiction

G Prompt for Data Annotation

In this section, we present an example prompt of the data annotation procedure for the three tasks.

G.1 Prompt for Data Annotation for Image Captioning Task

`System`
You are an annotator for image captioning tasks.
You will help create stylized image and its captions based on user requests.

`User` // *Image Decomposition*
Please generate a detailed prompt for DALL-E3 model to replicate the given image. [Original Image]
`Assistant`
Create an image of a man preparing food outside an industrial-style workspace. The man is wearing a flat cap and a dark short-sleeve shirt and is standing at a brown counter, chopping green onions on a cutting board. Surrounding him on the counter are various fresh vegetables, including green onions, leafy greens, a whole avocado, and a bowl of eggs. In the background, an open garage door reveals the interior of the workspace with tools, a workbench, and a bicycle leaning against the outside. The floor is concrete and the walls are decorated with hanging tools and shelves. The overall atmosphere should convey a casual, industrious vibe.

`User` // *Style Injection*
Please modify the generated prompt to change the style of the image to a cartoon drawing style.
`Assistant`
Create a cartoon-style image of a man preparing food outside an industrial-style workspace. The man is wearing a flat cap and a dark short-sleeve shirt and is standing at a brown counter, chopping green onions on a cutting board. Surrounding him on the counter are various fresh vegetables, including green onions, leafy greens, a whole avocado, and a bowl of eggs. In the background, an open garage door reveals the interior of the workspace with tools, a workbench, and a bicycle leaning against the outside. The floor is concrete and the walls are decorated with hanging tools and shelves. The overall atmosphere should convey a casual, industrious vibe, with cartoonish exaggerated features and vibrant colors.

`User` // *Image Verification*
Please verify if the image below is a cartoon drawing style image of the original image. The generated image should not exactly match the original image but should capture the essence of the original image. Start the response with 'Yes' or 'No'. [Generated Image]
`Assistant`
Yes, the provided image is a cartoon-style representation of the original image. The soccer player is depicted in a stylized, animated manner, with exaggerated features typical of cartoons. The attire, including the white short-sleeved jersey, blue shorts, long white socks, and white and orange cleats, closely matches the original image. The animated background with grass, orange cones, and a goal net also retains elements from the original setting, demonstrating a colorful and whimsical portrayal.

`User` // *Caption Paraphrasing*
Please generate five captions of the generated cartoon drawing style image. The captions should describe the image in different ways. Consider given captions below for reference.
1. a man cutting up vegetables on top of a food cart.
2. A man cutting up scallions at an outdoor table
3. A man chopping vegetables on a white board.
4. A man stands as he cuts up some vegetables.
5. A man stands with a knife and onions in front of a garage
`Assistant`
1. A cartoon-styled man slicing vegetables at an outdoor table near a garage.
2. An animated character chopping fresh scallions on a cutting board outside.
3. A cartoon figure preparing a meal by cutting vegetables on a white board in front of a workshop.
4. A man in cartoon form stands at an outdoor table, diligently cutting vegetables.
5. A cartoon man, standing with a knife and spring onions, prepares food outside a garage filled with tools.

G.2 Prompt for Data Annotation for Visual Question Answering Task

We omitted the image generation process as it is shared across three tasks.

`System`
You are an annotator for visual question answering tasks.
You will help create stylized image and its questions based on user requests.

// Image Generation Process

`User` // *Answer Verification*
Please verify if the question and answer pair below is correct for the generated cartoon drawing style image. Start the response with ‘Yes’ or ‘No’.
Question: Is the person wearing a hat?
`Assistant`
No, the question and answer pair is not correct. The person in the generated image is not wearing a hat.

`User` // *Answer Re-annotation*
Please answer the question below based on the generated cartoon drawing style image. Start the response with ‘Yes’ or ‘No’.
Question: Is the person wearing a hat?
`Assistant`
No, the person in the generated image is not wearing a hat.

`User` // *Question Paraphrasing*
Please paraphrase the question below for the generated cartoon drawing style image. The paraphrased question should have the same meaning as the original question but be rephrased in a different way. Only the question should be paraphrased.
Question: Is the person chopping green onions?
`Assistant`
Paraphrased Question: Is the individual slicing green onions?

G.3 Prompt for Data Annotation for Visual Entailment Task

`System`
You are an annotator for visual entailment tasks.
You will help create stylized image and its corresponding hypothesis based on user requests.

// Image Generation Process

`User` // *Label Verification*
Please verify if given hypothesis pair and its label is correct for the generated cartoon drawing style image. Start the response with 'Yes' or 'No'.
Hypothesis: The person is preparing ingredients for a meal in an outdoor kitchen setup.
Label: True
`Assistant`
No, the hypothesis "The person is preparing ingredients for a meal in an outdoor kitchen setup" is not entailed by the given image.

`User` // *Label Re-annotation*
Does the given hypothesis entail the image? Start the response with 'True', 'False', or 'Undetermined'.
Hypothesis: The person is preparing ingredients for a meal in an outdoor kitchen setup.
`Assistant`
Undetermined. It is unclear in the generated image if the person is preparing ingredients "in an outdoor kitchen setup."

`User` // *Hypothesis Paraphrasing*
Please paraphrase the hypothesis sentence below for the generated cartoon drawing style image. The paraphrased hypothesis should have the same meaning as the original sentence but be rephrased in a different way. Only the sentence should be paraphrased.
Hypothesis: The person is preparing ingredients for a meal in an outdoor kitchen setup.
`Assistant`
Paraphrased Hypothesis: The individual is getting ingredients ready for cooking in an outdoor kitchen setting.

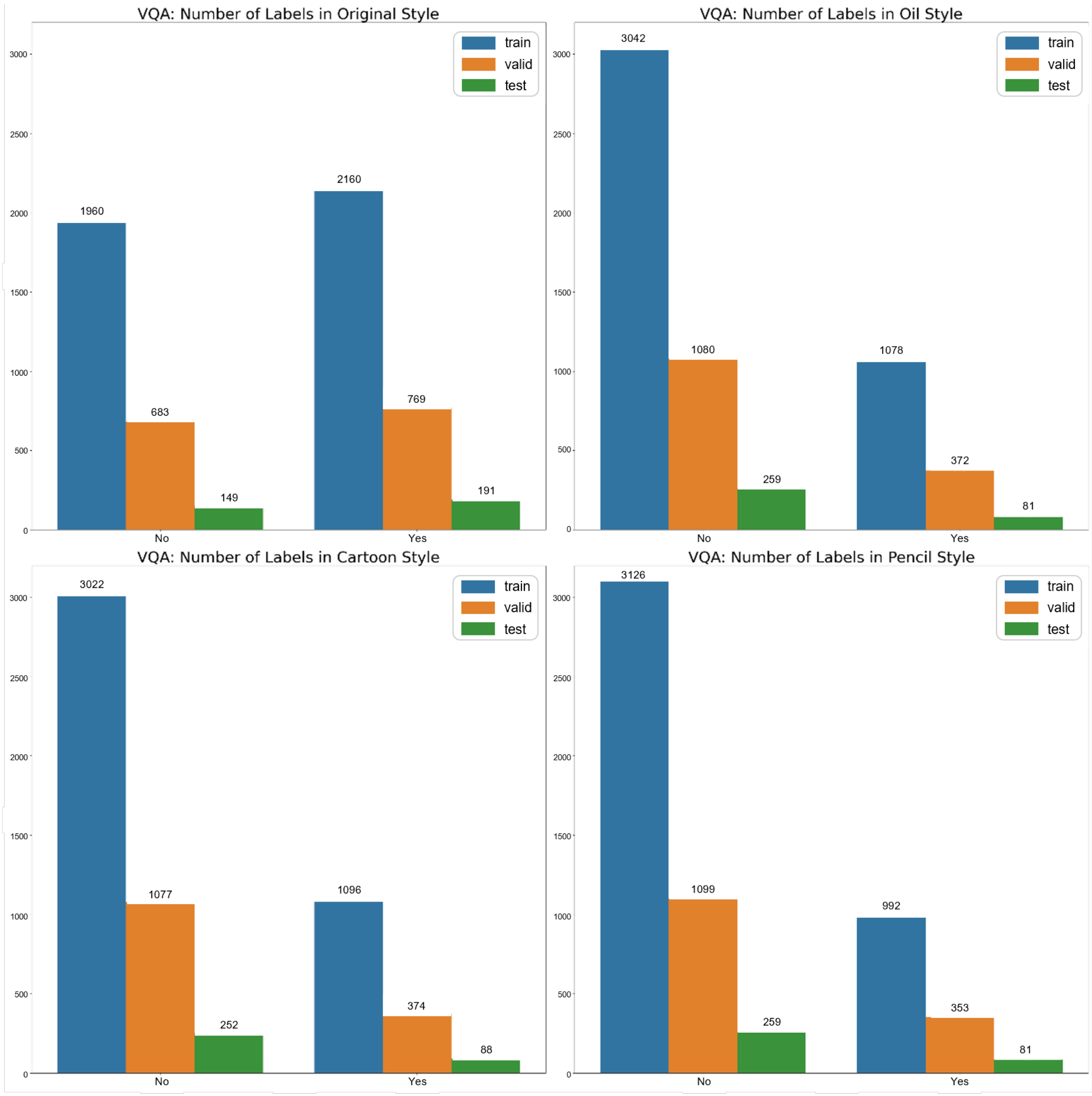


Figure 3: The label distribution of VOLDOGER-VQA for each split.

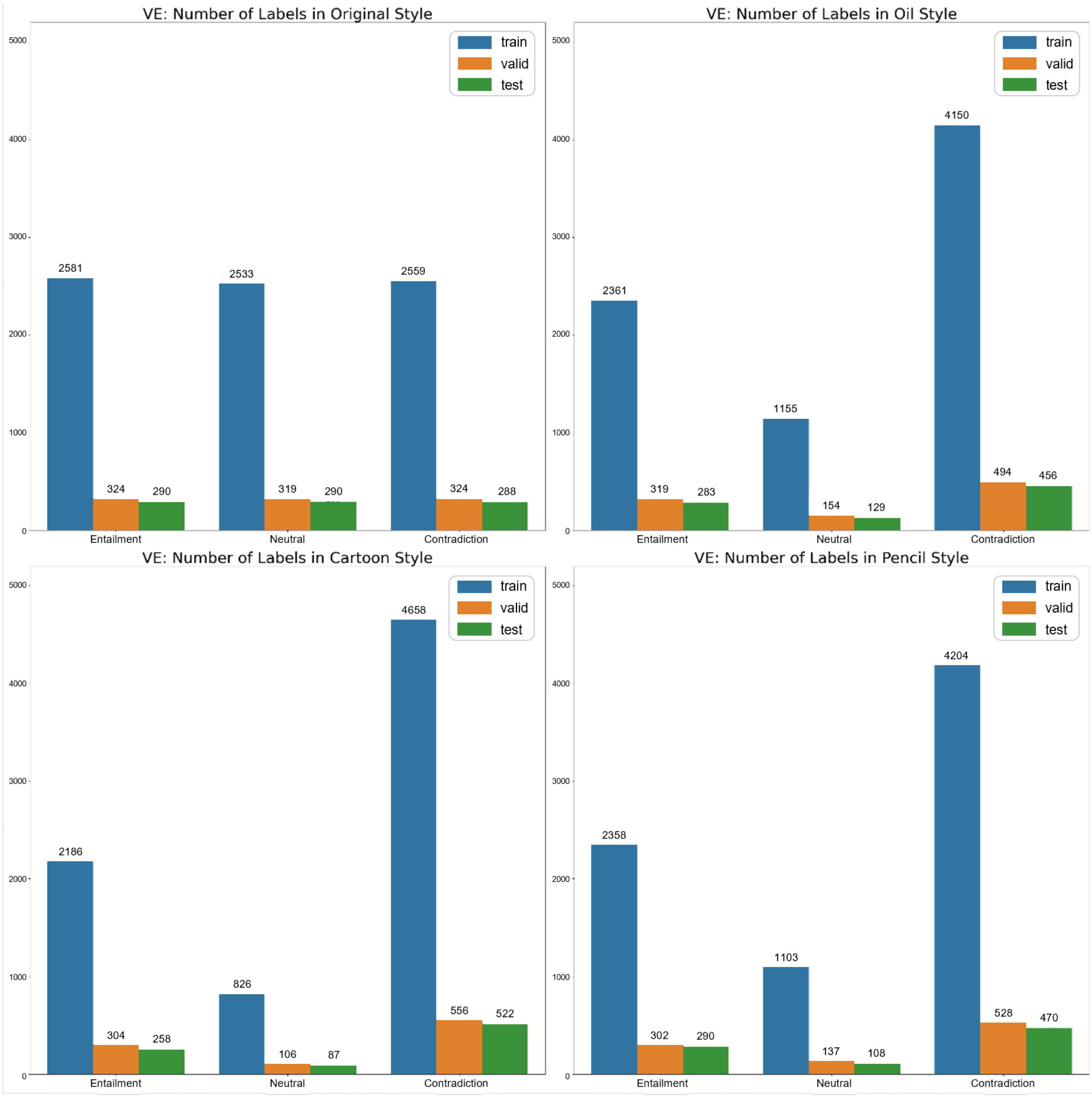


Figure 4: The label distribution of VOLDOGER-VE for each split.

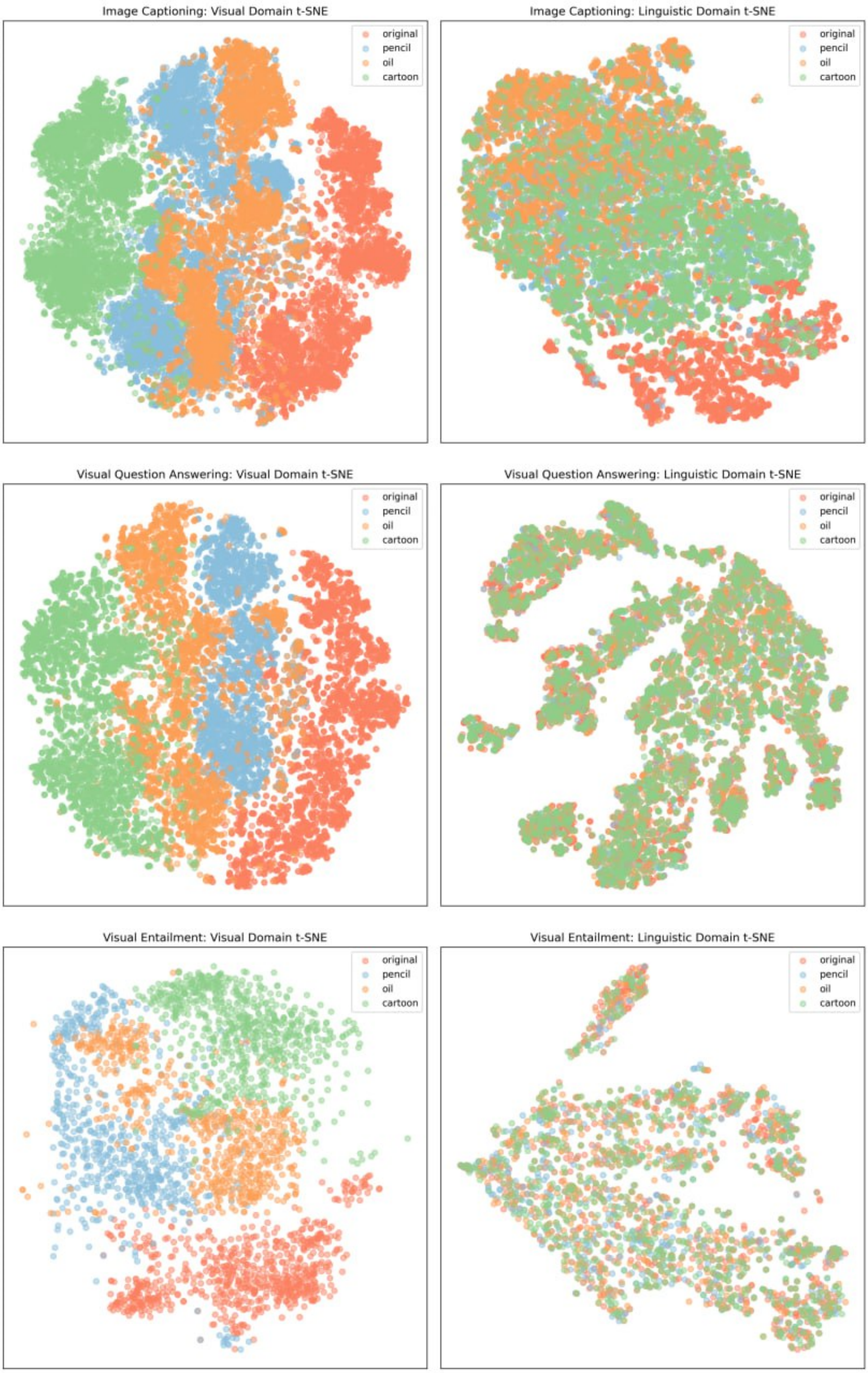


Figure 5: The t-SNE visualization result of each domain on three tasks.