

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

Supplementary Material

8. 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 [73]. We adopted open-source models such as BLIP-2 [42], PaliGemma [28], and LLaVA [45], as well as proprietary models such as GPT-4 [47], Gemini [27], and Claude 3 [2] for a comprehensive evaluation.

The results are listed in Tables 7, 8, and 9. 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 [40], where these open-source models may have used VQA-v2 data during their training process².

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 [12]. 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.

²Note 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 7. 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 [22].

9. Implementation Detail

This section presents implementation details to supplement the experimental setup described in Section 5.1. We primarily employed PyTorch [50] and Transformers [69] 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 details³.

9.1. 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 Trans-

³https://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 8. Experimental result demonstrating the zero-shot performance of multimodal large language models in VQA task.

formers library, respectively. For BLIP, we avoided directly 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 [38] 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 [57]. 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 [32] 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/ Vi-

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

cuna, 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 [59] 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.

9.2. 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 that do not support a dedicated system prompt, 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}?” This choice was made as open-source models such as PaliGemma do not support system prompts.

9.3. 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}”

9.4. Data Annotation

For data annotation, we used the GPT-4o [48] 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 14. In addition, we used DALL-E 3 [5] as the image generation model $\mathcal{G}$. Note that other image generation models such as Stable Diffusion [19] can also be used as $\mathcal{G}$ instead of DALL-E 3. The overall data annotation

procedure costs approximately USD 1,800.

10. Dataset Specification

In this section, we provide more detailed information on the VOLDOGER. Additionally, Figure 5 suggests the result t-SNE visualization [64] for each domain of three tasks, especially demonstrating visual domain gaps.

10.1. VOLDOGER-CAP

Table 10 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 10. The amount of images for each style in VOLDOGER-CAP.

10.2. VOLDOGER-VQA

Tables 11 and 12 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.

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 11. 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 12. 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.

10.3. VOLDOGER-VE

Tables 13 and 14 present the number of images, hypotheses, and the domain gap for each style in VOLDOGER-VE. Figure 4 presents the number of labels for each split.

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 13. Number of images and questions for each style in VOLDOGER-VE.

11. Ablation Study

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

	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 14. 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.

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 15. The result of ablation experiment that excludes answer verification process from our framework.

VQA and VE tasks.

11.1. 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.

11.2. 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 15. The findings suggest that directly assigning 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.

12. Further Experimental Result

12.1. Domain Shift of Model in Image Captioning

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

Table 16. Supplementary experimental result demonstrating domain shift on image captioning task.

12.2. Effectiveness of Domain Generalization in Image Captioning Task

ViT	Trained on R+C						Trained on R+P				
	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
<i>Real</i>	44.82	51.13	28.46	0.6838	-4.6268	<i>Real</i>	44.74	51.14	28.30	0.6828	-4.6261
<i>Cartoon</i>	41.95	41.64	23.35	0.6701	-4.8279	<i>Cartoon</i>	36.19	38.13	20.56	0.6410	-4.8279
<i>Pencil</i>	31.48	36.07	18.76	0.6275	-4.6130	<i>Pencil</i>	43.01	41.91	23.28	0.6519	-4.6116
<i>Oil</i>	30.09	32.99	17.23	0.6281	-4.7254	<i>Oil</i>	31.08	33.22	17.68	0.6287	-4.7277
Frozen CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
<i>Real</i>	50.10	54.53	30.61	0.6977	-4.6260	<i>Real</i>	49.20	54.48	30.54	0.6969	-4.6253
<i>Cartoon</i>	42.06	41.87	23.35	0.6718	-4.8264	<i>Cartoon</i>	35.69	37.50	20.61	0.6353	-4.8264
<i>Pencil</i>	30.64	35.48	18.75	0.6288	-4.6116	<i>Pencil</i>	43.51	41.17	23.02	0.6510	-4.6105
<i>Oil</i>	26.81	31.54	16.40	0.6170	-4.7253	<i>Oil</i>	29.12	32.14	17.66	0.6222	-4.7253
ViT w/ [57]	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
<i>Real</i>	39.97	48.21	27.84	0.6769	-4.6283	<i>Real</i>	46.96	35.16	29.50	0.6905	-4.6284
<i>Cartoon</i>	42.38	42.04	23.04	0.6649	-4.8264	<i>Cartoon</i>	36.65	38.56	21.61	0.6406	-4.8236
<i>Pencil</i>	31.72	36.38	19.17	0.6318	-4.6103	<i>Pencil</i>	42.05	40.79	23.23	0.6517	-4.6124
<i>Oil</i>	30.05	32.73	17.52	0.6306	-4.7235	<i>Oil</i>	31.59	34.17	17.70	0.6414	-4.7235
ViT	Trained on R+O						Trained on C+P				
	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
<i>Real</i>	45.49	52.24	28.98	0.6947	-4.6274	<i>Real</i>	20.65	33.31	22.14	0.5762	-4.6293
<i>Cartoon</i>	32.04	34.64	19.10	0.6265	-4.8269	<i>Cartoon</i>	42.54	42.18	23.24	0.6616	-4.8267
<i>Pencil</i>	32.73	34.08	18.66	0.6272	-4.6134	<i>Pencil</i>	43.12	41.14	23.62	0.6469	-4.6105
<i>Oil</i>	45.81	42.29	23.34	0.6749	-4.7254	<i>Oil</i>	34.14	34.79	18.97	0.6394	-4.7254
Frozen CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
<i>Real</i>	48.64	53.52	30.03	0.6938	-4.6292	<i>Real</i>	20.99	34.66	22.45	0.5771	-4.6290
<i>Cartoon</i>	33.35	34.42	19.86	0.6297	-4.8266	<i>Cartoon</i>	42.82	42.20	23.64	0.6733	-4.8264
<i>Pencil</i>	31.74	34.53	18.88	0.6323	-4.6122	<i>Pencil</i>	43.09	41.64	23.72	0.6519	-4.6103
<i>Oil</i>	45.68	42.04	23.05	0.6711	-4.7353	<i>Oil</i>	34.50	34.91	19.12	0.6353	-4.7255
ViT w/ [57]	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
<i>Real</i>	44.03	50.75	28.22	0.6811	-4.6259	<i>Real</i>	23.89	35.50	22.89	0.6353	-4.6253
<i>Cartoon</i>	34.42	35.32	20.75	0.6366	-4.8264	<i>Cartoon</i>	41.71	40.62	23.12	0.6445	-4.8265
<i>Pencil</i>	34.99	34.85	19.92	0.6324	-4.6112	<i>Pencil</i>	42.51	41.55	23.37	0.6483	-4.6111
<i>Oil</i>	44.70	41.43	22.65	0.6721	-4.7224	<i>Oil</i>	35.88	35.42	19.20	0.6409	-4.7253
ViT	Trained on C+O						Trained on R+P				
	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
<i>Real</i>	20.66	32.13	22.10	0.5763	-4.6265	<i>Real</i>	20.24	32.22	21.67	0.5778	-4.6256
<i>Cartoon</i>	43.20	42.48	24.13	0.6781	-4.8266	<i>Cartoon</i>	32.24	34.71	20.13	0.6285	-4.8273
<i>Pencil</i>	35.17	36.11	20.14	0.6316	-4.6109	<i>Pencil</i>	44.23	42.28	24.02	0.6547	-4.6106
<i>Oil</i>	46.74	42.35	23.69	0.6789	-4.7264	<i>Oil</i>	47.17	43.04	23.81	0.6761	-4.7253
Frozen CLIP	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
<i>Real</i>	20.62	32.51	22.62	0.5815	-4.6257	<i>Real</i>	20.31	31.95	21.08	0.5782	-4.6267
<i>Cartoon</i>	43.69	42.85	23.83	0.6816	-4.8264	<i>Cartoon</i>	34.49	37.22	21.08	0.6380	-4.8268
<i>Pencil</i>	35.60	36.19	20.49	0.6345	-4.6159	<i>Pencil</i>	44.42	42.02	23.95	0.6549	-4.6128
<i>Oil</i>	46.06	42.08	24.22	0.6801	-4.7245	<i>Oil</i>	46.94	43.37	24.13	0.6788	-4.7256
ViT w/ [57]	BLEU	ROUGE	METEOR	BERTS.	BARTS.		BLEU	ROUGE	METEOR	BERTS.	BARTS.
<i>Real</i>	22.57	32.86	22.76	0.5829	-4.6253	<i>Real</i>	21.77	32.98	22.40	0.5839	-4.6253
<i>Cartoon</i>	42.76	42.04	23.81	0.6751	-4.8268	<i>Cartoon</i>	36.58	38.23	21.93	0.6402	-4.8266
<i>Pencil</i>	36.83	36.45	20.75	0.6429	-4.6103	<i>Pencil</i>	43.04	41.32	23.39	0.6498	-4.6133
<i>Oil</i>	46.88	42.77	23.58	0.6766	-4.7253	<i>Oil</i>	47.01	42.80	23.67	0.6737	-4.7261

Table 17. 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.

12.3. Experimental Results in Visual Question Answering Task

VQA	Trained on Cartoon Drawing				Trained on Pencil Drawing				Trained on Oil Painting			
	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>
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			
	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>
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/ [57]	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			
	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>
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/ [57]	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 18. Supplementary experimental result demonstrating the domain shift and effectiveness of domain generalization technique on VQA task.

12.4. Experimental Results in Visual Entailment Task

VE	Trained on Cartoon Drawing				Trained on Pencil Drawing				Trained on Oil Painting			
	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>
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			
	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>
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/ [57]	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			
	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>	<i>Real</i>	<i>Cartoon</i>	<i>Pencil</i>	<i>Oil</i>
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/ [57]	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 19. Supplementary experimental result demonstrating the domain shift and effectiveness of domain generalization technique on VE task.

13. Additional Examples of Annotated Data

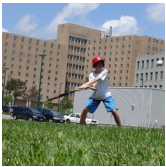
13.1. Additional Examples on Image Prompt

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.

13.1.1. 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.

13.1.2. Pencil Drawing Style Images with Prompts

Original Image with <i>p_ori</i>	Stylized Image with <i>p sty</i>
 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.

13.1.3. Oil Painting Style Images with Prompts

Original Image with <i>Pori</i>	Stylized Image with <i>Pay</i>
 A photograph of two men standing in a park. The man on the left is wearing a yellow t-shirt with bold text and light-colored shorts. The man on the right is wearing a white shirt, white shorts, and a white cap, with his arm around the other man's shoulder. In the background, a third man in a white shirt and black shorts is walking away while looking down at something in his hands, possibly 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. Some buildings or structures are faintly visible further back.	 An oil painting of the same scene. The style is characterized by warm, textured brushstrokes and rich, vibrant colors typical of oil paintings. 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, white shorts, and a white cap, with his arm around the other man's shoulder. Both are portrayed with friendly expressions. In the background, a third man in a white shirt and black shorts, also wearing a cap, is walking away while looking down at something in his hands, such as a notebook or a phone. The ground is 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.
 A photograph 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.	 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.
 A photograph 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 polished wooden flooring is also visible, capturing the energy and intensity of the match.	 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.

13.2. 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 brushwork. • 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.

13.3. 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 girl's right arm in an awkward position? • Answer: Yes	 • Question: Is the girl's right arm positioned awkwardly? • Answer: No

13.4. 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

14. Prompt for Data Annotation

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

14.1. 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.
-

14.2. 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?

14.3. 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.

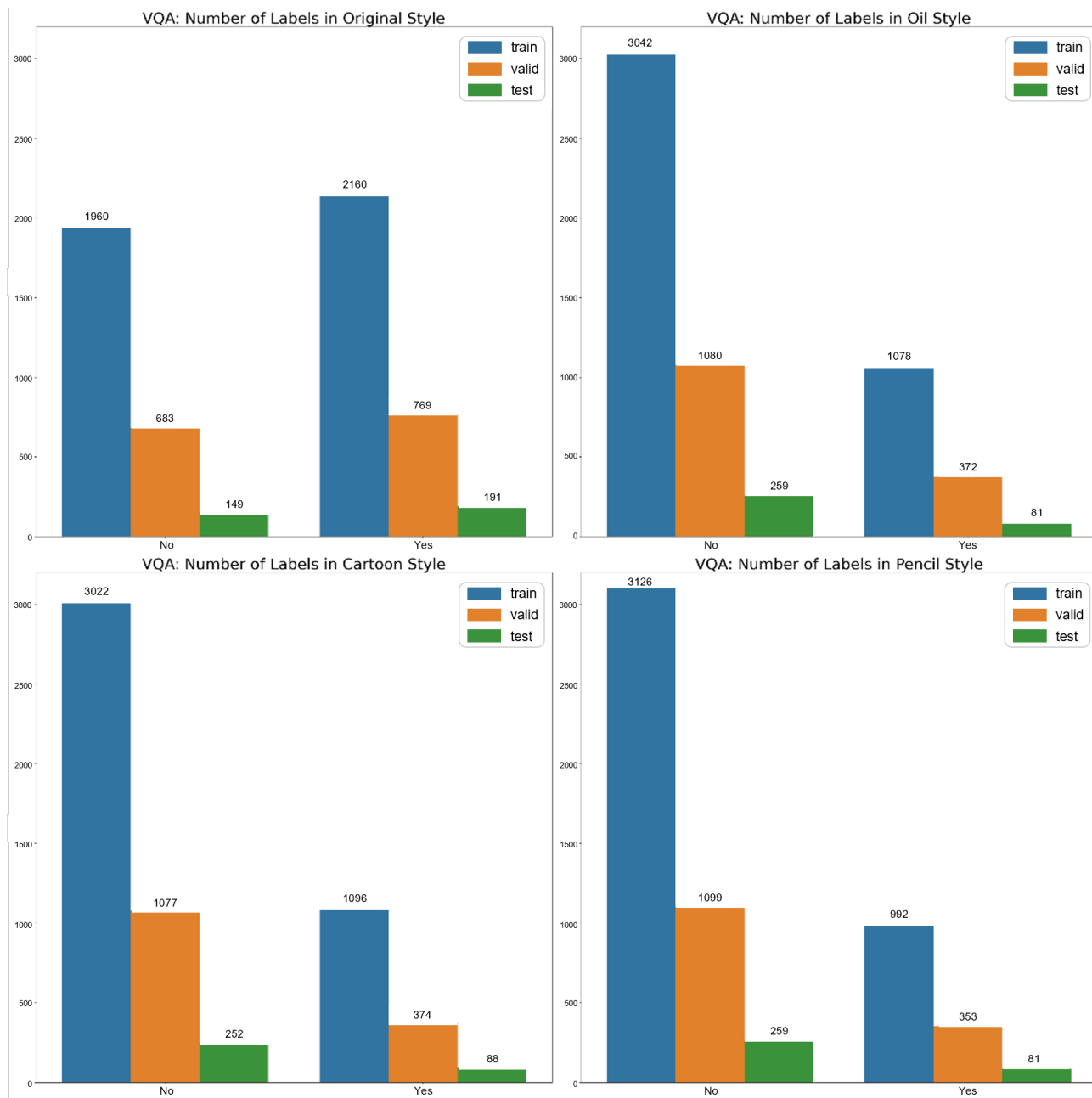


Figure 3. The label distribution of VOLDGER-VQA for each split.

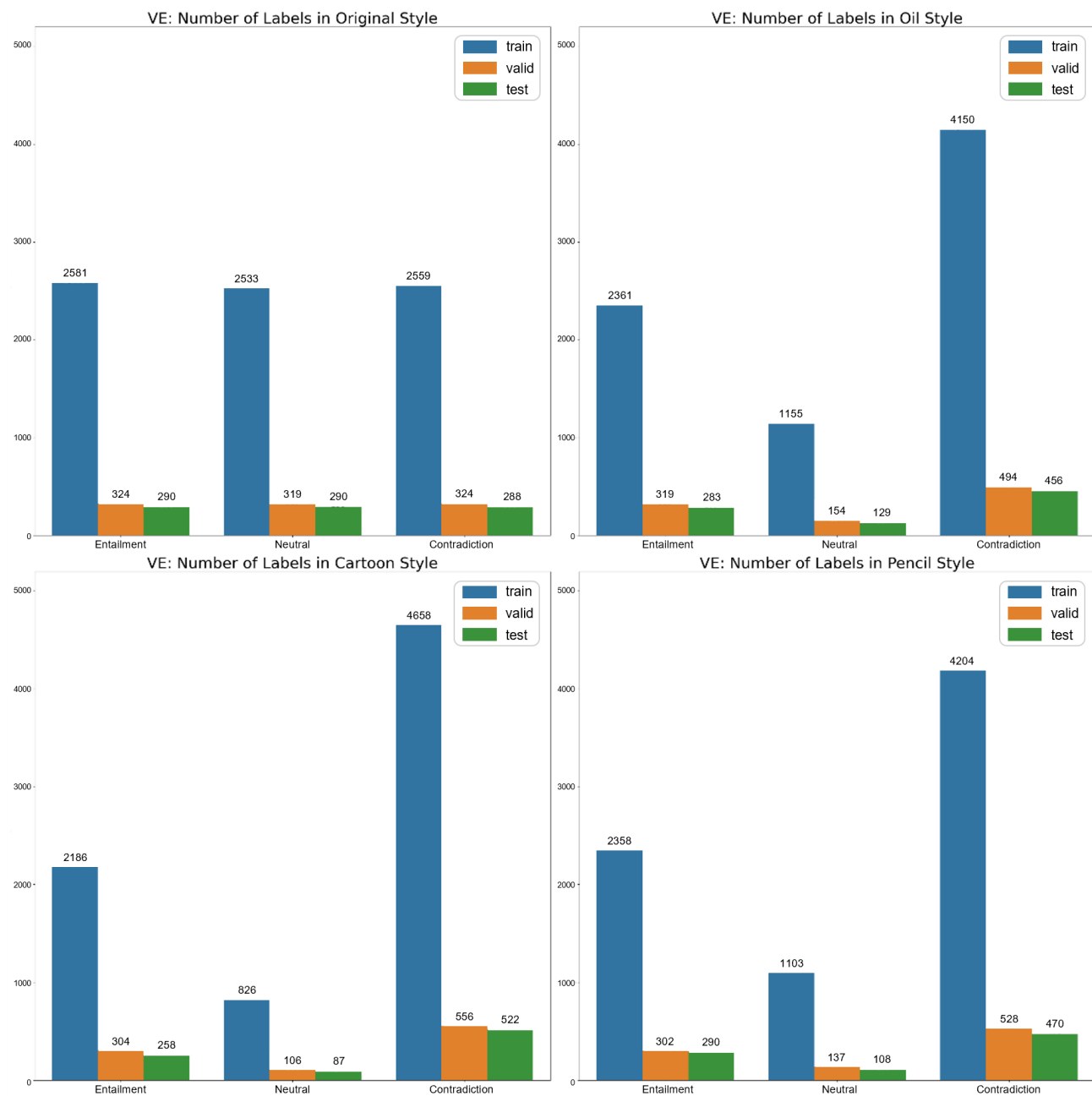


Figure 4. The label distribution of VOLDOGER-VE for each split.

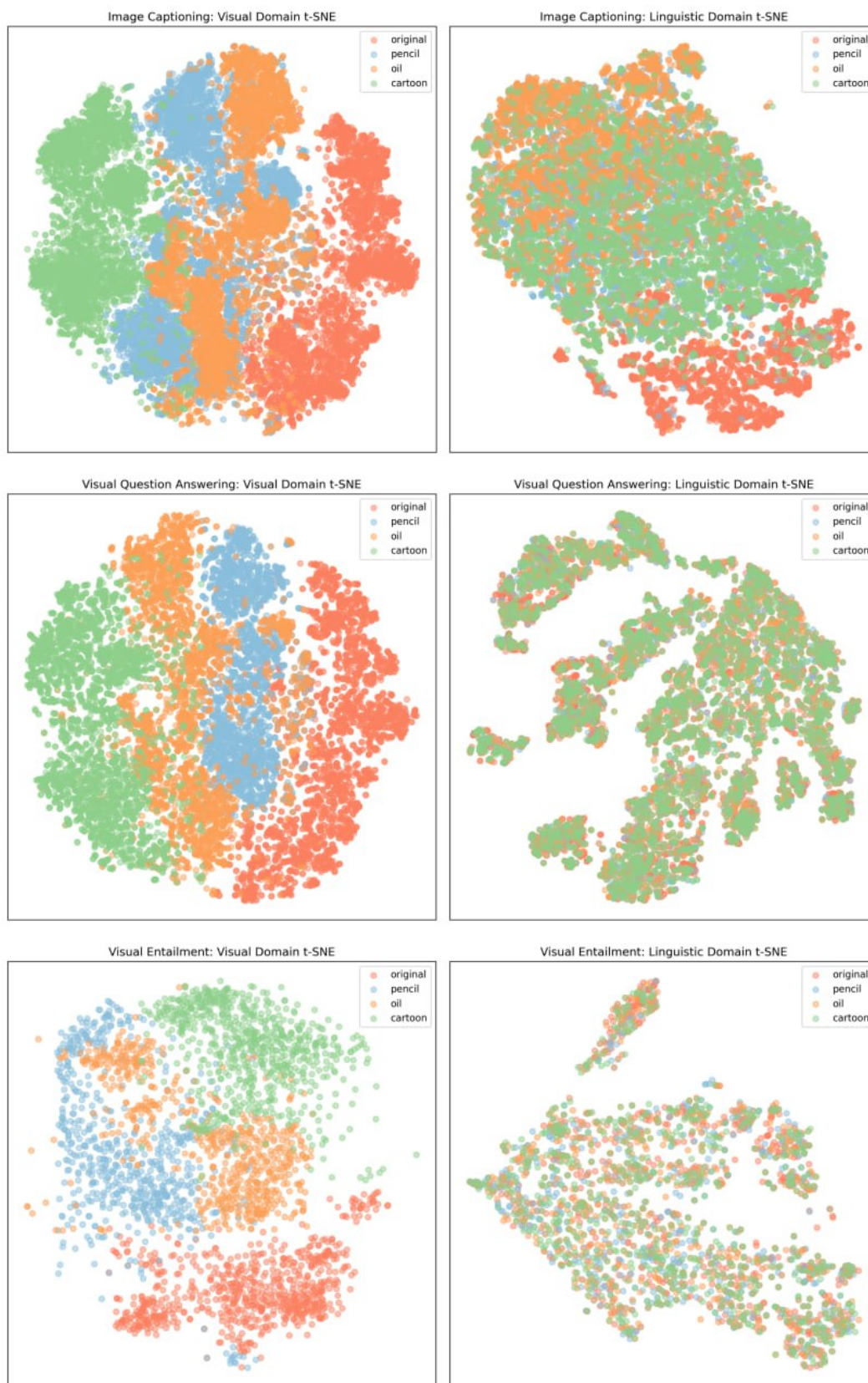


Figure 5. The t-SNE visualization result of each domain on three tasks.