A Related Works

A.1 Instruction Tuning

Instruction tuning (Zhang et al., 2023; Ouyang et al., 2022; Chung et al., 2024; Zheng et al., 2023) is a strategy commonly adopted in modern LLM training to enhance model generalization by exposing models to various prompts. In the realm of multimodal large language models (MLLMs), visual instruction tuning (Liu et al., 2024) has significantly improved their instruction-following capabilities when processing multimodal data. This process typically involves two stages: the first stage trains an adapter between the visual encoder and the LLM using image captioning data; in the second stage, the LLM and the adapter are jointly trained with instruction-following data that encompasses multiple tasks in a question-answer format. While previous MLLMs have primarily focused on text generation, recent research is exploring the use of LLMs for representation learning. Specifically, E5-Mistral (Wang et al., 2023) leverages LLMs as embedding models by training them on various retrieval tasks specified by instructions. E5-V (Jiang et al., 2024) extends this approach to multimodal domains; however, its training remains based on pure text pairs, and the full potential of MLLMs for multimodal embeddings is not fully realized. In this paper, we propose a novel approach to train an instruction-aware model that generates multimodal embeddings through two stages: embedding alignment and instruction contrastive learning.

A.2 Composed Image Retrieval

Composed Image Retrieval (CIR) involves finding images related to a source image under a specified condition, typically provided as a modifier text. This task has practical applications in e-commerce, recommendation systems, and more. Due to the difficulty of acquiring specific datasets for various CIR tasks, recent research has focused on Zero-Shot CIR (ZS-CIR). Previous methods primarily represent the reference image as specific tokens and concatenate them with text tokens for retrieval (Saito et al., 2023; Karthik et al., 2023; Tang et al., 2024; Suo et al., 2024; Agnolucci et al., 2024; Gu et al.). With the advent of Multimodal Large Language Models (MLLMs), researchers have begun incorporating LLMs into this domain. For instance, CIReVL (Karthik et al., 2023) leverages two MLLMs: one for generating image captions and another for combining captions with modifier texts for retrieval. FROMAGe (Koh et al., 2023) and MCL (Li et al.) explore using LLMs for embeddings, but the LLMs are mainly used as text encoders. Despite the rapid development of MLLMs exhibiting strong generalization, instruction-following, and zero-shot capabilities in multimodal data, their applications to CIR tasks are rarely explored. In this paper, we leverage MLLMs as embedding models for CIR tasks, enabling direct encoding of images and modifier texts within a single model.

B Triplet Data Generation

B.1 Data Processing

We utilize GPT-4o (Achiam et al., 2023) to process and generate triplet data. Given an image and its caption, we use the caption as a prompt to GPT, which then derives the modifier text and the modified caption. The detailed prompt structure is shown in Figure 1. Specifically, the prompt is divided into three parts: task definition, requirements, and few-shot examples.

Our data generation process differs from MCL (Li et al.) in several aspects. First, we leverage GPT-4o (Achiam et al., 2023) instead of LLAMA2 (Touvron et al., 2023), allowing for more generalizable and creative content generation. Second, GPT-4o has a larger context window, enabling us to incorporate more complex techniques within the prompt. Unlike MCL, which directly presents the output modifier text and corresponding caption in few-shot examples, we divide the generation process into several steps using the Chain of Thought method (Wei et al., 2022). We instruct GPT to first identify key points in the example caption, then selectively alter some of them as modifications, and finally derive the modified caption. This step-by-step generation ensures that the generated modifier text and corresponding caption are reasonable and closely related to the original caption. *At the time the major work of this paper is finished, the MCL dataset has not been released. We will deffer the comparison between two datasets in the future work.*

Our pipeline differs from the training set derivation in (Vaze et al., 2023). While they use text scene graphs to identify subjects, predicates, and objects, their modifier instruction is generated by simply replacing one element with another concept from the dataset, leading to limited creativity and diversity.

I am creating a multi-modal dataset for Composed Image Retrieval (CIR). The goal is to generate pairs of source and target images, along with a modification instruction that describes how to transform the source image into the target image.

Your Task:
1. Input: I will provide you with a source image caption.
2. Instruction Generation: Brainstorm a modification instruction based on the source caption. This instruction should be a clear, concise description of a plausible change that can be applied to the source image.
3. Modified Caption Generation: Apply the modification instruction to the source caption to create a modified caption that describes the target image after the change.
4. You should output the modification instruction and modified caption only.

Requirements:
1. The modification instruction should focus on a single, significant change (e.g., changing an object's color, altering the setting, modifying an action).
2. The modified caption should reflect only the changes specified in the instruction while keeping the rest of the description consistent with the source caption.
3. Ensure that the instruction and modified caption are coherent and plausible.

Example #1:
Input:
Source Caption: A Husky is lying on the grass.
Brainstorming:
The caption contains an object husky, an action lying, and a background grass. One plausible change is altering the action of the dog from lying to running. The modified caption then becomes: a husky is running on the grass.
Output:
Modification Instruction: The dog is running.
Modified Caption: A husky is running on the grass.

Example #2: ……

Input:
Source Caption: a very typical bus station

Brainstorming:
The caption describes a location, "a very typical bus station". One significant change could be altering the time of day, which affects the lighting and activity at the location. Transitioning from day to night can introduce new elements like artificial lighting and perhaps a quieter atmosphere.

Output:
Modification Instruction: Change the time of day to night.
Modified Caption: A very typical bus station at night.

Figure 1: We prompt GPT-4o to generate triplet data from CC3M. Our prompt consists of three parts: the first part (orange) defines the task we aim to complete; the second part (blue and purple) specifies the details and requirements of the task; and the third part (black) provides examples for triplet generation, where the modifier text is brainstormed step by step. The key concepts in the captioned are identified and subsequently selected concepts are altered. The modified caption is derived accordingly. Finally, we provide the input (red). GPT then outputs the modifier text and the corresponding caption based on the query caption (green).

B.2 Data Details

After filtering invalid images and failed prompts, we acquire the CC3M-Instruct dataset with 2M triplets. Triplet examples are shown in Figure 2.

C Prompt Templates

Templates for training are shown in Table 1.

C.1 Templates for Training

Task	Instruction Template
Image Modification	`<Image>` The image is conditioned on the following prompt: {modifier text}, summarize the image and the prompt to retrieve a description of the image changed by the condition:
	`<Image>` Given the image conditioned by the prompt: {modifier text}, condense the essence of the image and the prompt into a single word to fetch a description of the altered image:
	`<Image>` Using the prompt to condition the image: {modifier text}, provide one word that encapsulates the overall concept of the conditioned image to retrieve its description:
	`<Image>` Based on the image influenced by this prompt: {modifier text}, distill the description of the conditioned image and the prompt into one word to access the altered description:
	`<Image>` With the image modified according to the prompt: {modifier text}, summarize both the image and the prompt to obtain a description of the conditioned image:
	`<Image>` Condition the image with this condition: {modifier text}. Summarize the result:
	`<Image>` Using this prompt: {modifier text}, describe the conditioned image:
	`<Image>` Apply the prompt: {modifier text} to the image. Provide one word for the conditioned image:
	`<Image>` Given this prompt: {modifier text}, condense the conditioned image into one word:
	`<Image>` {modifier text}:
Image Summary	`<Image>` Summary:
	`<Image>` Caption:
	`<Image>` Summarize the image for retrieval:
	`<Image>` A short image caption:
	`<Image>` A short image description:
	`<Image>` Provide a description of what is presented in the photo:
	`<Image>` Please provide a short depiction of the picture:
	`<Image>` Using language, provide a short account of the image:
	`<Image>` Use a word to illustrate what is happening in the picture:
Caption Summary	`<Caption>` Summary:
	`<Caption>` Summarize the caption for retrieval:
	`<Caption>` A shorter description is:
	`<Caption>` Shorter caption:
	`<Caption>` “”

Table 1: Instruction templates for different tasks. In **Image Modification**, the modifier text combined with the selected template serves as the formatted prompt. **Image** and **Caption Summary** instruct the model to generate a global representation for images or captions.

C.2 Templates for Zero-shot Inference

CIRR & CIRCO

Image Captioning

<Image> Describe this image in one word:

Image Modification

<Image> Modify this image with {modifier text}, describe the modified image in one word:

FashionIQ

Image Captioning

<Image> Describe this {data type in fashioniq} in one word based on its style:

Image Modification

<Image> Modify the style of this {data type in fashioniq} based on {modifier text}. describe this modified {data type in fashioniq} in one word based on its style:

GeneCIS

Image Captioning

<Image> Summarize the image for retrieval:

Image Modification

<Image> Describe the image in one word with a specific focus on the attribute {specific attribute}:

<Image> Describe the image in one word with a specific change of the attribute {specific attribute}:

<Image> Describe the image in one word with a specific focus on the object {specific object}:

<Image> Describe the image in one word with a specific change of on the object {specific object}:

D TRAINING DETAILS

D.1 MLLM TRAINING

We use the code and data from xtuner/llava-phi-3-mini-hf (Contributors, 2023) to train a variant of LLaVA-Phi. *Note that the goal of this step is solely to make our experiments consistent with the baselines.* Section **??** has demonstrated that our training strategy can be directly applied to existing MLLMs. The checkpoint of the variant LLaVA-Phi will also be released for reproducibility. MLLM training and model details are provided as follows.

Config	Value
Visual Encoder	openai/clip-vit-large-patch14
Image Resolution	224x224
Language Model	microsoft/Phi-3.5-mini-instruct
Adapter	MLP
Pretraining Strategy	Frozen LLM, Frozen ViT
Fine-tuning Strategy	Full LLM, Full ViT
Pretrain Dataset	ShareGPT4V-PT (1246K) (Chen et al., 2023)
Fine-tune Dataset	InternVL-SFT (1268K) (Chen et al., 2024)
Pretrain Epoch	1
Fine-tune Epoch	2

Table 2: Configurations of Training LLaVA-Phi

	xtuner/llava-phi-3-mini-hf	microsoft/Phi-3.5-vision-instruct	E5-V
Size	4.14B	4.15B	8.35B

Table 3: Number of parameters of different models

D.2 INSTRUCTCIR TRAINING

Detailed training configs are shown in Table 4.

Training Config	Value
DeepSpeed	ZeRO-2
LoRA R	64
LoRA Alpha	16
Model Max Length	512
Precision	FP16
Epochs for both stages	1
Batch Size Per GPU in Stage 1	48
Batch Size Per GPU in Stage 2	64
Gradient Accumulation Steps	1
Learning Rate	2E-05
Weight Decay	0
Warm Up Ratio	0.03
LR Scheduler Type	Cosine

Table 4: Configurations of Training InstructCIR.

E MORE EXPERIMENT RESULTS

Table 5, 6, 7 demonstrate the complete results of InstructCIR that is trained with LLaVA-Pretrain (Liu et al., 2024) only in the first training stage.

Method	CIRR						CIRCO			
	R@1	R@5	R@10	R_s@1	R_s@2	R_s@3	mAP@5	mAP@10	mAP@25	mAP@50
InstructCIR$_{lp}$	35.08	**65.25**	76.53	67.52	84.13	92.08	22.19	23.62	26.01	27.20
InstructCIR$_{full}$	**35.18**	65.12	**77.61**	**67.54**	**84.77**	**93.61**	**22.32**	**23.80**	**26.25**	**27.32**

Table 5: **Comparison of Zero-Shot CIR Models on CIRCO and CIRR Test Sets.** InstructCIR$_{lp}$ refers to InstructCIR that is trained with LLaVA-Pretrain only in the first training stage. InstructCIR$_{full}$ is trained with both LLaVA-Pretrain and FOIL in the first training stage.

Method	Shirt		Dress		Toptee		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
InstructCIR$_{lp}$	29.85	49.98	25.04	45.60	31.74	53.26	28.90	49.61
InstructCIR$_{full}$	**30.96**	**50.10**	**25.11**	**46.18**	**32.32**	**54.22**	**29.46**	**50.16**

Table 6: **Comparison of Zero-Shot CIR Models on FashionIQ**. InstructCIR$_{lp}$ refers to Instruct-CIR that is trained with LLaVA-Pretrain only in the first training stage. InstructCIR$_{full}$ is trained with both LLaVA-Pretrain and FOIL in the first training stage.

Method	Focus Attribute			Change Attribute			Focus Object			Change Object			Average
	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1
InstructCIR$_{lp}$	20.35	33.35	45.05	15.39	28.39	37.69	16.58	26.69	**37.19**	**17.14**	27.86	**38.62**	17.37
InstructCIR$_{full}$	**21.25**	**34.55**	**46.85**	**16.15**	**28.74**	**39.73**	**17.55**	**28.01**	36.94	17.04	**28.98**	37.70	**18.00**

Table 7: **Comparison of Zero-Shot CIR Models on GeneCIS**. InstructCIR$_{lp}$ refers to InstructCIR that is trained with LLaVA-Pretrain only in the first training stage. InstructCIR$_{full}$ is trained with both LLaVA-Pretrain and FOIL in the first training stage.

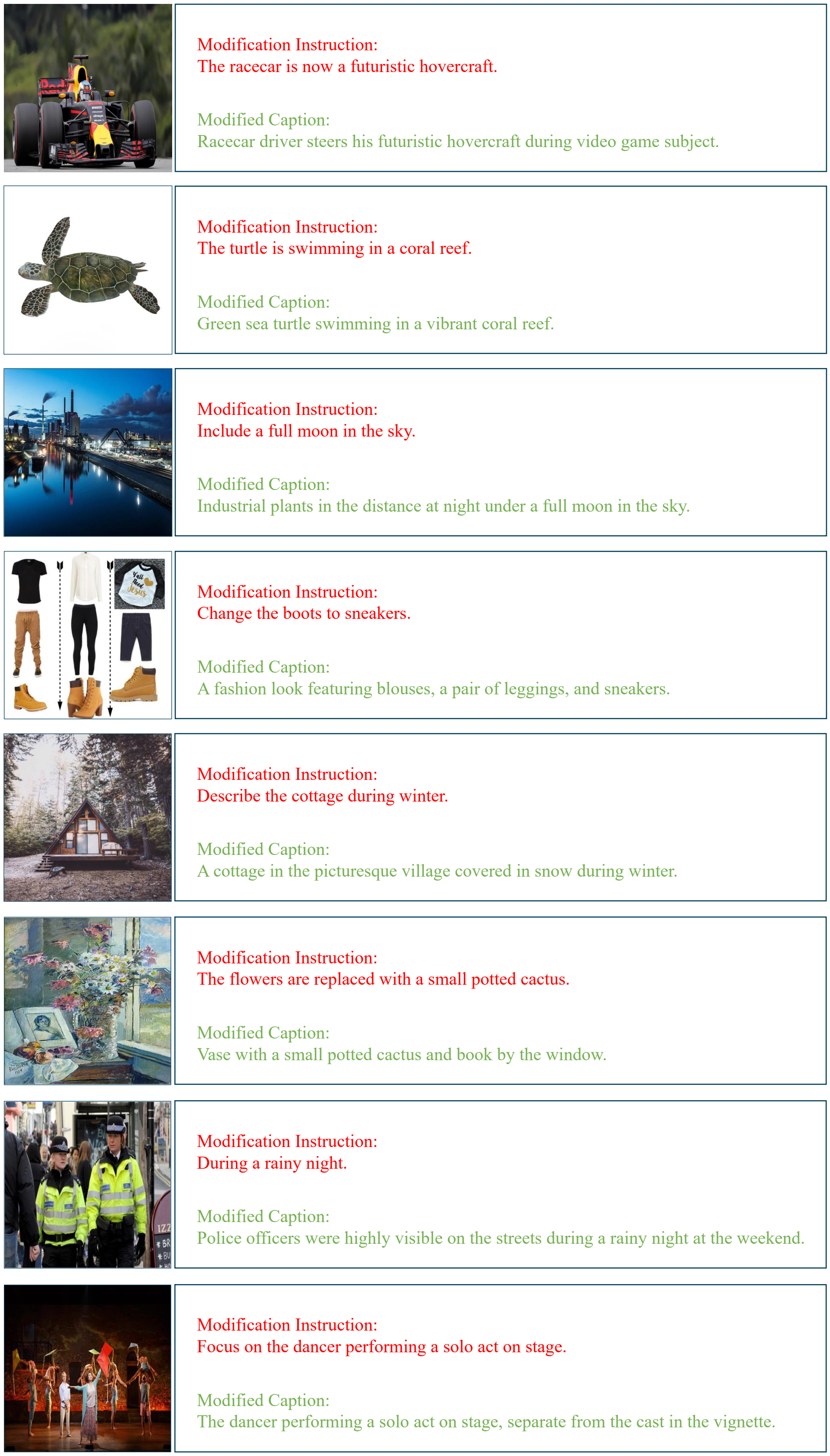

Figure 2: Triplet Examples from CC3M-Instruct

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Lorenzo Agnolucci, Alberto Baldrati, Marco Bertini, and Alberto Del Bimbo. isearle: Improving textual inversion for zero-shot composed image retrieval. *arXiv preprint arXiv:2405.02951*, 2024.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. `https://github.com/InternLM/xtuner`, 2023.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval–appendix–.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. 2023.

Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan Kankanhalli. Improving context understanding in multimodal large language models via multimodal composition learning. In *Forty-first International Conference on Machine Learning*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19305–19314, 2023.

Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26951–26962, 2024.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5180–5188, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6862–6872, 2023.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.