

MA-CIR: A Multimodal Arithmetic Benchmark for Composed Image Retrieval

Supplementary Material

A1. More Examples

We provide additional examples of hard negatives for the color and left-right categories in Fig. A1 and Fig. A2, respectively.

We include examples for the replacement and addition arithmetic types (omitted in Fig. 4) in Fig. A3 and Fig. A4. As mentioned in Section 4, we omit the *addition* cases for the spatial reasoning and size categories, as these cases rely on semantic relations defined between two objects. In these cases, as shown in Fig. A4, the conditioning text (such as “is ahead of” or “behind” for spatial reasoning, and “the bigger” or “the smaller” for size) cannot be effectively represented in an image with an empty background, as the relative location and size of objects cannot be clearly defined.

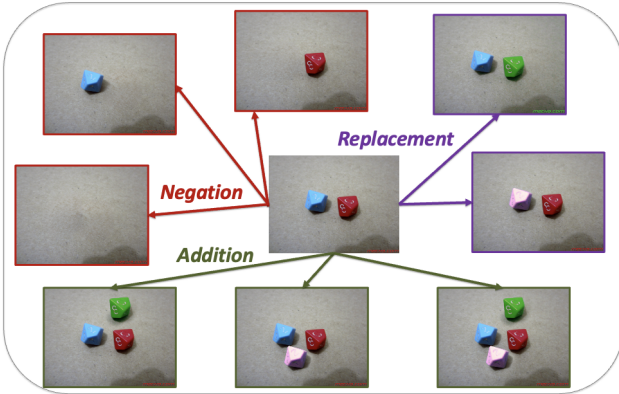


Figure A1. Examples of hard negatives in MA-CIR with “color” category

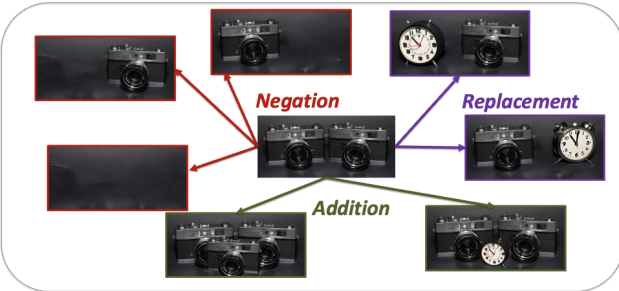


Figure A2. Examples of hard negatives in MA-CIR with “left-right” category

A1.1. Examples of human evaluation.

As shown in Fig. A5, we incorporate the instructions and examples used in human evaluation for MA-CIR. Each evaluator assesses whether the reference image, conditioning

text, and target image are appropriately composed based on the specified composition and condition types and whether the generated image is reasonable to recognize the corresponding composition or condition types.

A2. Additional Details and Results

A2.1. Results on backbone size and type

We assess the impact of backbone size and type with various methods. As verified in Sec. 6, simply using different backbones has only a marginal impact. The entire results for Fig. 5 and Fig. 6 are also included here.

Table A1. Impact of fine-tuned backbone types. (Neg-CLIP and FSC-CLIP)

Method	Size	LR	TB	SR	S	AC	C	OR	NO	Avg
LinCIR [6]		19.4	22.6	21.7	17.3	23.6	21.8	16.1	16.4	19.9
LinCIR [6] + Neg [19]	ViT-B/32	24.6	25.1	10.6	19.6	22.7	23.5	20.7	23.1	21.2
LinCIR [6] + FSC [14]		25.0	24.7	20.5	24.4	28.9	23.8	17.5	20.9	23.2
LinCIR [6]	ViT-L/14	21.4	27.6	18.6	20.2	25.2	26.9	21.8	25.0	23.3
LinCIR [6] + FSC [14]		24.6	30.9	28.0	21.4	31.4	21.4	14.7	20.2	24.1
Slerp [7]		23.0	18.5	21.7	20.8	28.1	22.4	14.0	16.4	20.6
Slerp [7] + Neg [19]	ViT-B/32	23.4	28.4	23.6	22.0	28.9	21.4	13.0	17.9	22.3
Slerp [7] + FSC [14]		24.6	28.0	29.8	28.0	31.8	19.7	14.7	17.5	24.3
Slerp [7]	ViT-L/14	23.0	18.5	21.7	20.8	28.1	22.4	14.0	16.4	20.6
Slerp [7] + FSC [14]		31.8	32.1	24.2	23.8	29.3	21.8	15.1	18.7	24.6

Table A2. Impact of backbone sizes (ViT-B/32, ViT-L/14, ViT-G/14).

Method	Size	LR	TB	SR	S	AC	C	OR	NO	Avg
LinCIR [6]	ViT-B/32	19.4	22.6	21.7	17.3	23.6	21.8	16.1	16.4	19.9
	ViT-B/16	23.0	25.1	21.7	25.6	25.6	23.1	19.0	19.8	22.9
	ViT-L/14	21.4	27.6	18.6	20.2	25.2	26.9	21.8	25.0	23.3
	ViT-H/14	29.0	25.5	16.2	22.0	24.0	20.8	20.4	23.1	22.6
	ViT-G/14	21.4	23.1	18.0	17.3	31.0	22.5	26.3	23.9	22.9
Slerp [7]	ViT-B/32	16.7	25.1	21.7	23.2	31.0	20.1	16.5	15.3	21.2
	ViT-B/16	22.6	20.6	21.7	22.0	26.9	18.7	15.4	14.9	20.7
	ViT-L/14	23.0	18.5	21.7	20.8	28.1	22.5	14.0	16.4	20.6
	ViT-H/14	26.6	23.9	22.4	21.4	29.3	22.5	19.0	16.4	22.6
	ViT-G/14	23.0	21.4	28.6	22.0	28.5	19.7	16.1	13.1	21.6
CIReVL [9]	ViT-B/32	28.6	17.3	23.0	37.5	24.4	38.4	29.8	35.1	29.3
	ViT-L/14	30.4	24.7	20.7	30.6	21.4	32.3	19.3	23.1	25.3
	ViT-G/14	32.1	21.8	25.5	39.9	30.2	42.9	31.9	34.3	32.3
MagicLens [20]	ViT-B/32	36.1	32.1	27.3	30.4	26.9	33.3	10.9	19.4	27.0
	ViT-L/14	35.7	34.6	30.4	28.0	31.0	38.4	14.0	19.8	29.0
SPRC [1]	ViT-L/14	28.2	29.6	25.5	32.7	31.0	31.6	15.8	22.4	27.1
	ViT-G/14	25.0	35.4	28.6	28.6	34.7	27.9	13.0	24.3	27.2

A2.2. Results on additional baselines (MCL [10] and CoVR [18])

In Tab. A3, we additionally evaluate two baselines trained on synthetic CIR triplets: CoVR fine-tunes BLIP backbones,

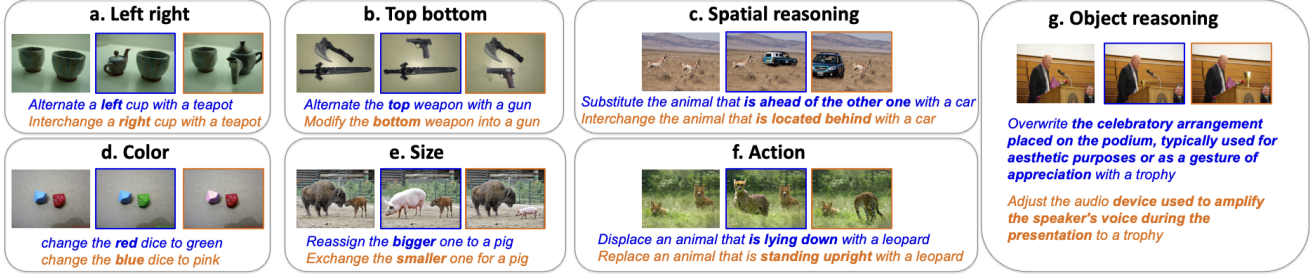


Figure A3. Examples of “replacement” arithmetic type for each category are included.

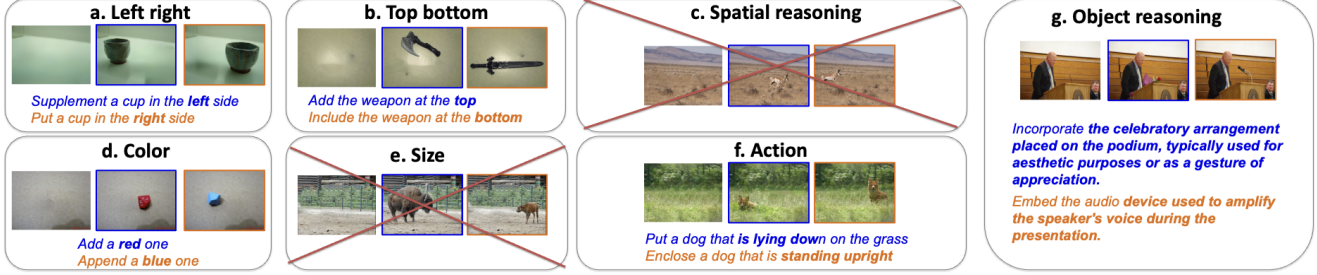


Figure A4. Examples of “addition” arithmetic type for each category are included.

Welcome to the Human Evaluation Task


In this task, you will evaluate triplets consisting of a reference image, conditioning text, and a target image.

Each triplet represents a relationship where: **reference image + conditioning text = target image**.

Your role is to assess the appropriateness of these triplets, ensuring that the conditioning text aligns with the reference/target images and that the images themselves are suitable.


Correct Example:


Reference Image:



→

Target Image:






Conditioning Text: Change the table that has an item placed on it with a chair.


Incorrect Example:


Reference Image:



→

Target Image:





Conditioning Text: Get rid of the right giraffe.

Click the "Start" button when you are ready to begin.

Start

Figure A5. Instructions for human evaluation.

while MCL trains adapter modules atop a frozen LLM. Consistent with our main results, both show limited performance on MA-CIR, suggesting a possible need for more carefully designed synthetic datasets and LLM adaptation.

Method	LR	TB	SR	S	A	C	OR	NO	Avg
CoVR [18]	18.7	25.9	18.0	19.0	34.3	24.8	16.5	17.4	21.8
MCL [10]	20.6	23.0	24.2	22.0	32.6	31.3	14.7	19.4	23.5

Table A3. Additional baselines for MA-CIR

Table A4. More results on E5-V and “Ours”

Method	MA-CIR									CIRCO mAP@5	CIRR R@1	FashionIQ	
	LR	TB	SR	S	A	C	OR	NO	Avg			R@10	R@50
E5-V [8] (official repo)	40.1	30.9	26.1	26.8	40.1	41.8	31.6	34.7	34.0	20.5	33.9	31.8	53.8
E5-V [8] (reproduced)	38.4	28.8	21.3	25.7	35.3	43.2	32.3	34.3	32.4	17.6	30.9	28.4	49.1
Ours	47.7	37.8	41.6	56.9	46.6	58.6	41.7	57.2	48.5	26.5	36.8	29.6	50.8

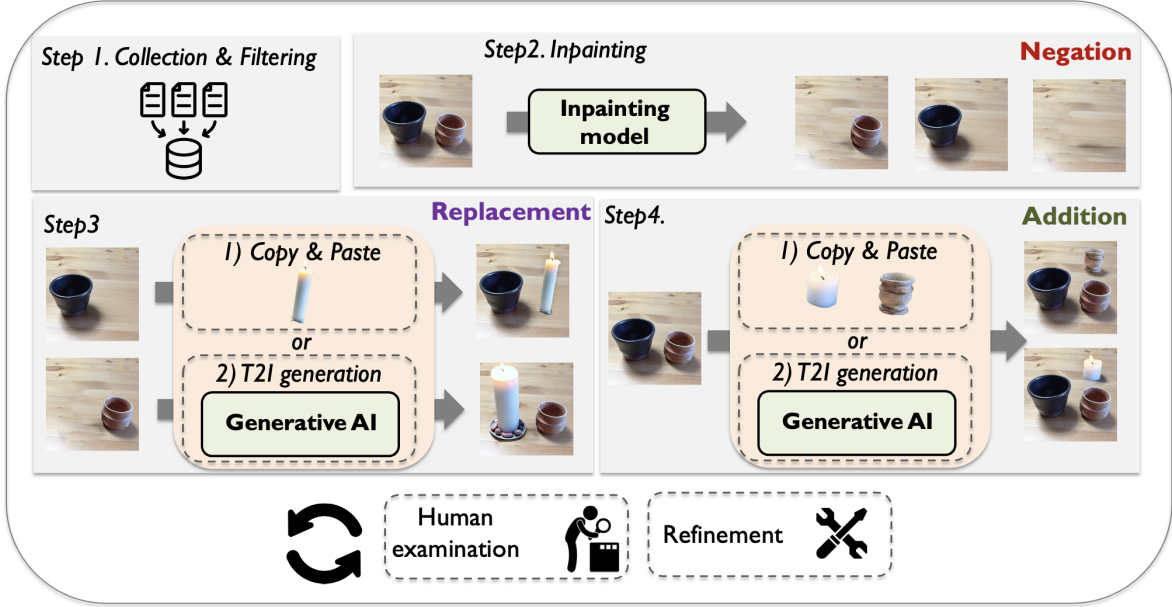


Figure A6. Generation pipeline.

A2.3. Implementation details on our adaptation.

We use LLaVA-NeXT-8B [12], built on LLaMA-3 8B, with a frozen ViT-L/14 as the visual encoder. The LLM of LLaVA is fine-tuned for up to 2000 iterations with a batch size of 64 using a single A100 Gpu. The best validation model for CIRR R@1 score (val split) is chosen following [6]. We employ QLoRA and gradient checkpointing with DeepSpeed ZeRO-2 for efficient training. The training prompt follows the format: “[T_r] that modifies this image with [T_c]. Describe the modified image in one word: ”. The prompt for generating our text triplets is described in Fig. A7. Since the training environment of the original E5-V model may differ from ours, we reproduce E5-V under our setup. Due to the high variance observed in reproducing E5-V, we report the average results over twelve runs for both “E5-V (reproduced)” and “Ours”. “E5-V (official repo)” refers to a single-run result. As shown in Tab. A4, “Ours” consistently and significantly outperforms both “E5-V (reproduced)” and “E5-V (official repo)”. While “Ours” slightly underperforms the official E5-V on FashionIQ, it surpasses the fairly reproduced E5-V under identical settings, demonstrating the effectiveness of our simple remedy.

A2.4. Results for each arithmetic type in Tab. 1

We provide detailed results for each arithmetic type listed in Tab. 1. All hard negatives are included in the evaluation as in Tab. 1; however, the results are measured separately for each arithmetic type. Note that the difference with results in Tab. 2 is that all hard negatives are included in Tab. A5. As explained in the Sec. 6, existing methods struggle more with negation and replacement arithmetic types compared to the addition type.

A2.5. Few-shot learning.

We apply few-shot fine-tuning to LinCIR, SEARLE, and Bi-Blip4CIR on MA-CIR, updating all network parameters, including CLIP backbones and projection module (or fusion module). We use a batch-based contrastive loss function, where the bi-modal query embedding q_i and its corresponding target image embedding v_i are paired for contrastive learning:

$$\mathcal{L}_{few} = \frac{1}{B} \sum_{k=1}^B -\log \frac{e^{(c(q_i, v_i)/\tau)}}{\sum_{j=1}^B e^{(c(q_i, v_j)/\tau)}} \quad (1)$$

Table A5. **Detailed results for Tab. 1.** N, R, A denotes an arithmetic types: *Negation*, *Replacement*, *Addition*, respectively. “Avg” value in Tab. 1 may differ slightly from the one shown here, as the average is computed over the metrics available in each respective table.

Method	Left Right			Top Bottom			Spatial Reasoning			Size			Action			Color			Object Reasoning			Naive Object			Avg
	N	R	A	N	R	A	N	R	A	N	R	A	N	R	A	N	R	A	N	R	A	N	R	A	
(a) Zero-shot																									
Text-only	3.6	23.8	6.0	5.8	22.8	11.5	2.5	33.8	-	0.0	17.9	-	8.6	29.6	28.8	0.0	7.3	14.7	0.0	3.6	9.7	0.0	6.2	10.3	11.2
Image-only	7.1	0.0	7.1	9.3	0.0	5.1	11.1	2.5	-	9.5	1.2	-	7.4	0.0	10.0	7.8	17.4	22.1	15.9	1.2	24.8	15.6	1.2	24.7	9.1
Image+Text	7.1	2.4	10.7	10.5	1.3	9.0	12.4	2.5	-	11.9	3.6	-	12.4	1.2	16.2	4.4	24.8	34.7	9.1	2.4	42.5	7.8	2.5	42.3	12.3
Slerp [7]	19.1	34.5	15.5	11.6	26.6	18.0	13.6	30.0	-	7.1	34.5	-	13.6	30.9	40.0	1.1	30.3	33.7	1.1	8.3	28.3	2.2	13.6	32.0	20.3
Slerp+TAT [7]	20.2	42.9	20.2	16.3	32.9	32.1	13.6	37.5	-	8.3	35.7	-	18.5	28.4	45.0	2.2	26.6	47.4	0.0	9.5	25.7	2.2	9.9	41.2	23.5
Pic2Word [16]	9.5	31.0	21.4	17.4	31.7	20.5	8.6	25.0	-	7.1	27.4	-	11.1	25.9	40.0	2.2	38.5	49.5	6.8	13.1	41.6	5.6	13.6	46.4	22.5
SEARLE [3]	20.2	25.0	22.6	23.3	29.1	26.9	17.3	21.2	-	15.5	21.4	-	14.8	23.5	45.0	2.2	26.6	45.3	2.3	14.3	33.6	5.6	13.6	40.2	22.2
LinCIR [6]	14.3	26.2	23.8	25.6	25.3	32.1	12.4	25.0	-	13.1	27.4	-	8.6	24.7	42.5	3.3	27.5	48.4	1.1	14.3	43.4	5.6	14.8	51.5	23.2
FTI4CIR [11]	15.5	23.8	23.8	24.4	34.2	28.2	8.6	27.5	-	9.5	25.0	-	18.5	22.2	56.2	7.8	33.0	55.8	2.3	10.7	50.4	2.2	8.6	50.5	24.5
Context-I2W [17]	8.3	25.0	20.2	18.6	32.9	20.5	12.4	28.8	-	6.0	25.0	-	13.6	28.4	43.8	3.3	31.2	46.3	2.3	14.3	31.9	3.3	12.4	44.3	21.5
LinCIR+RTD [4]	14.3	26.2	29.8	17.4	34.2	33.3	11.1	33.8	-	15.5	29.8	-	9.9	24.7	53.8	6.7	44.0	55.8	3.4	21.4	41.6	12.2	23.5	53.6	27.1
CIReVL [9]	32.1	42.9	13.1	24.4	30.4	5.1	22.2	31.2	-	31.0	46.4	-	22.2	24.7	31.2	33.3	46.8	38.9	17.0	26.2	33.6	24.4	40.7	37.1	29.8
E5-V [8]	45.6	32.9	36.7	21.6	30.1	35.6	20.3	22.3	-	18.2	33.2	-	18.3	18.4	69.5	17.8	46.6	63.4	11.6	8.8	65.9	19.0	13.2	66.1	32.5
Ours	70.7	37.3	34.9	40.6	38.0	34.6	43.7	39.5	-	63.3	50.5	-	44.1	30.0	65.9	55.4	58.0	62.5	42.1	19.8	57.5	62.3	36.7	69.4	48.0
(b) Supervised CIR																									
Combiner [2]	4.8	2.4	9.5	7.0	2.5	14.1	11.1	2.5	-	7.1	4.8	-	11.1	1.2	16.2	4.4	24.8	29.5	12.5	2.4	37.2	12.2	2.5	33.0	11.5
Bi-Blip4CIR [13]	19.0	35.7	22.6	15.1	26.6	20.5	12.3	16.2	-	17.9	27.4	-	17.3	17.3	46.3	20.0	36.7	44.2	5.7	14.3	39.8	14.4	16.0	41.2	23.9
SPRC [15]	33.3	35.7	15.5	33.7	38.0	16.7	17.3	33.8	-	22.6	42.9	-	12.4	21.0	60.0	12.2	33.0	48.4	6.8	15.5	23.0	11.1	17.3	37.1	26.7
(c) Synthetic data-based CIR																									
Compodiff [5]	13.1	22.6	16.7	11.6	34.2	19.2	19.8	30.0	-	9.5	31.0	-	9.9	28.4	32.5	3.3	45.0	28.4	8.0	19.1	31.0	6.7	22.2	41.2	22.0
MagicLens [20]	44.1	41.7	21.4	34.9	41.8	26.9	29.6	31.2	-	21.4	34.5	-	12.4	30.9	50.0	15.6	44.0	53.7	3.4	7.1	27.4	10.0	7.4	39.2	28.6

Table A6. **Few-shot learning with SEARLE, LinCIR, and Bi-Blip4CIR.**

Method	LR	TB	SR	S	AC	C	OR	NO	Avg
SEARLE [3]	22.6	26.3	19.3	18.5	27.7	25.2	18.3	20.5	22.3
+Few-shot FT	22.6	25.1	18.0	20.2	22.7	28.9	23.5	25.4	23.3
LinCIR [6]	21.4	27.6	18.6	20.2	25.2	26.9	21.8	20.5	23.3
+Few-shot FT	17.5	24.7	14.9	17.3	25.6	25.9	20.4	22.0	21.0
Bi-Blip4CIR [13]	25.8	20.6	14.3	22.6	26.9	34.0	21.8	24.6	23.8
+Few-shot FT	27.0	22.6	16.8	27.4	26.4	32.3	20.0	23.5	24.5

where $c(\cdot, \cdot)$ denotes the cosine similarity, B is the batch size, and τ is a temperature. Here, the bi-modal query embedding is obtained using the prompt “a photo of [\$] that [T_c]”, and [$\$$] is the projected text token embedding of I_r with the projection module. Both the reference and target image embeddings are obtained using the same visual encoder. Although LinCIR was originally trained on text-only data, we fine-tune it using MA-CIR triplets (which include images), following the same protocol as other methods. The small-scale training MA-CIR sets (10-64 samples), which do not overlap with the evaluation set, are used for fine-tuning. We follow the default training setup for Bi-BLIP4CIR. For LinCIR and SEARLE, we set the learning rate to $1e - 5$

when updating all modules, including the text encoder, image encoder, and projection module. The maximum number of training iterations is set to 50 for all methods. In Table A6, all results show marginal effectiveness, highlighting the challenging nature of our benchmarks, which cannot be addressed by simple methods with small-scale training data.

A3. Qualitative Results

In Figs. A8 to A10, we present the top-5 predictions of the baseline methods: Slerp [7], LinCIR [6], SPRC [1], MagicLens [20], E5-V [8], and our adaptation method. In the figure, we observe that the zero-shot CIR methods (Slerp, LinCIR) struggle to handle both arithmetic operations (e.g., *negation* and *replacement*) and complex semantic relationships. In contrast, MagicLens, E5-V, and our adaptation method demonstrate better performance in capturing arithmetic operations, with relevant images ranked higher. In particular, E5-V and our adaptation method, which leverage multi-modal large language models (MLLM), exhibit a superior understanding of both arithmetic and semantic types compared to other methods.


```

keywords(replace)=["replace", "swap", "substitute", "exchange", "switch", "alter", "modify", "transform", "convert", "adjust", "revise", "interchange", "update", "overwrite", "reassign", "rearrange", "change"]

keywords(remove)=["remove", "delete", "erase", "eliminate", "exclude", "detach", "discard", "withdraw", "subtract", "omit", "extract", "clear", "cancel", "abolish", "dissolve", "wipe", "without"]

keywords(add)=["add", "insert", "include", "append", "attach", "place", "embed", "introduce", "put", "merge", "integrate", "combine", "incorporate", "join", "implant", "allocate", "deploy", "adjoin", "supplement", "with"]

message = {

"role": "user",

"content": f"""You are required to generate one example considering the given examples. The editing attributes should also be diverse. Make sure the examples are clear, concise, comprehensive. Describe the first caption in "{ref_caption}" like "input", the second caption in "OUTPUT_DESCRIPTION" like "output", both "INPUT_DESCRIPTION" and "OUTPUT_DESCRIPTION" should be independent complete sentences, and the description that edits the first caption to the second caption in "EDIT_DESCRIPTION" like "edit".

Use the following keywords to make the "EDIT_DESCRIPTION" as natural as possible: {keyword}.
The "EDIT_DESCRIPTION" should be natural and only contain one of the keywords mentioned above. Make sure to use each keyword at least once across multiple examples, and do not use "and".
The output should be a list of JSON format as such:

{{ "input": "{ref_caption}",
"edit": "EDIT_DESCRIPTION",
"output": "OUTPUT_DESCRIPTION" }}.
Do not output anything else, an example should have complete keys "input", "edit", and "output"."""

}

```

Figure A7. **Prompt for text triplet generation by LLMs used in Section 5.** It enables independent modification of subject and attribute information.

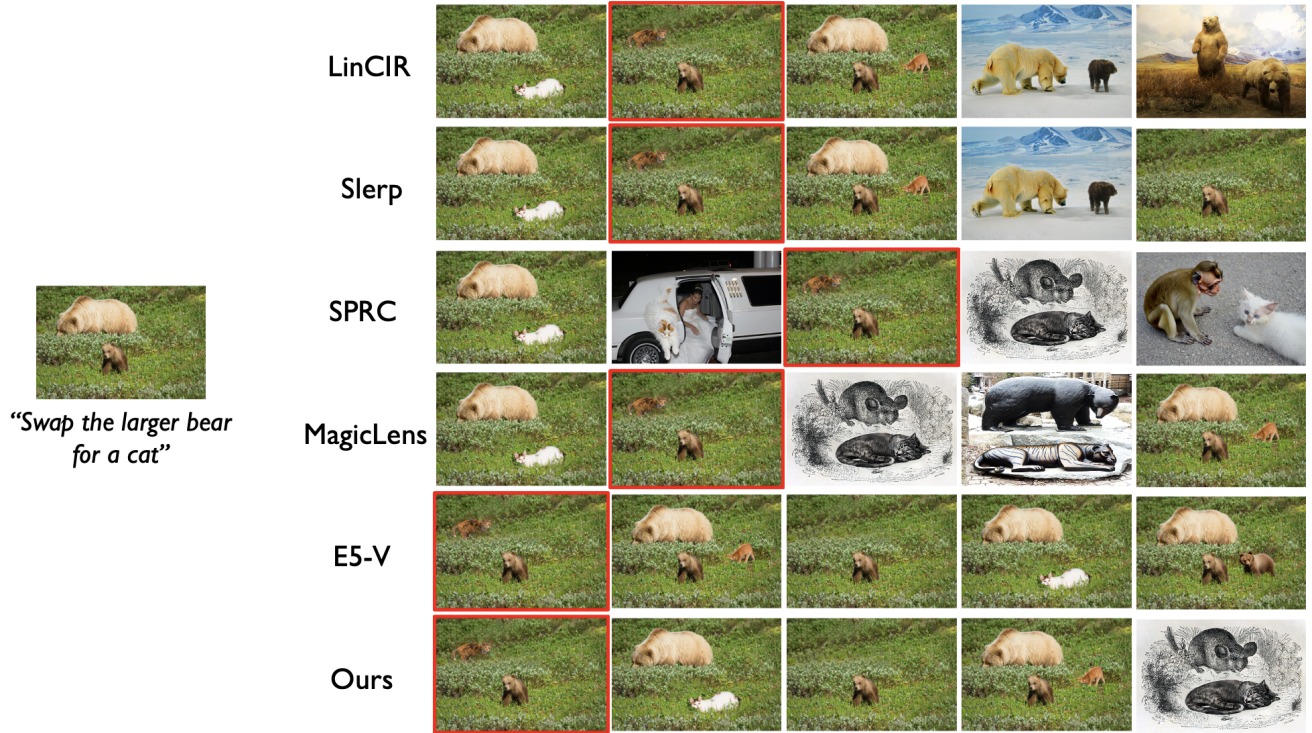


Figure A8. **Retrieval results in MA-CIR with "replacement" and "size" type.**

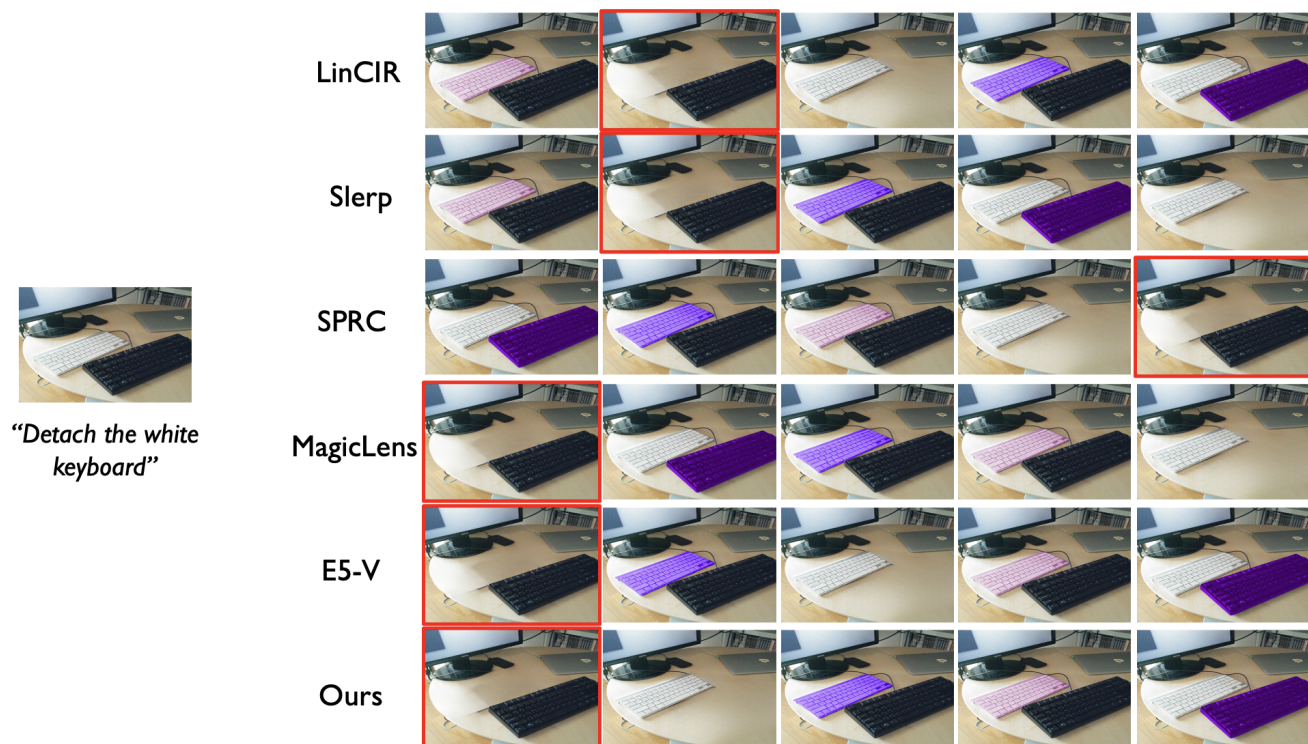


Figure A9. Retrieval results in MA-CIR with “negation” and “color” type.

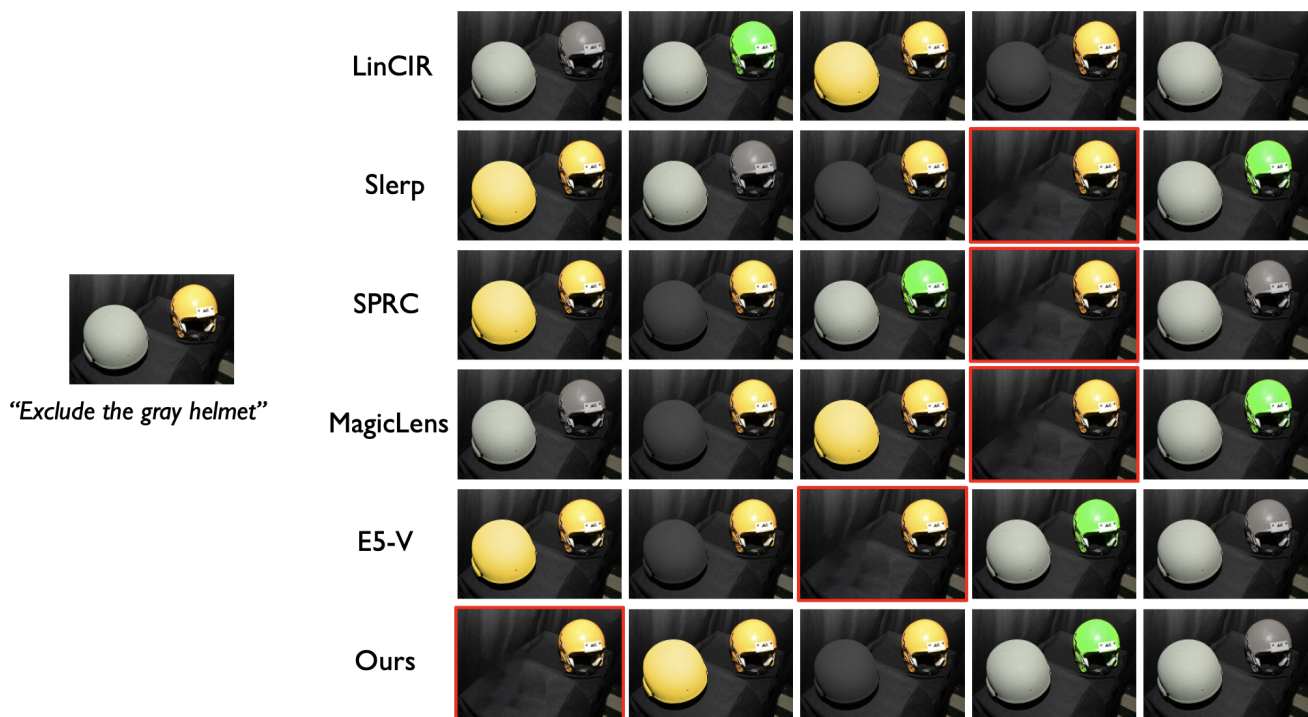


Figure A10. Retrieval results in MA-CIR with “negation” and “color” type.

A4. Impact of artifacts from generative models

Each MA-CIR triplet is constructed through careful manual and iterative refinement as shown in Fig. A5 and Fig. A6. Although minor artifacts may be present, human evaluators consistently judged the MA-CIR triplets to be appropriate for capturing the intended compositional and arithmetic types. Moreover, to assess whether the models evaluated in MA-CIR exhibit similar behavior in the absence of generative artifacts, we construct a small-scale dataset (≈ 100 queries) that avoids the use of generative or inpainting models. Instead, we capture real photographs of physical objects arranged in compositions resembling those in MA-CIR. The results closely align with the original MA-CIR trends, with average deviations under 3% from the scores in Tab. 1, suggesting that minor generative artifacts do not significantly affect the core findings.

A5. Discussions

MA-CIR does not cover all semantic categories (*e.g.*, object count, or viewpoint changes). However, as discussed in the Introduction, MA-CIR is designed to complement existing CIR benchmarks by focusing on overlooked areas (*e.g.*, negation and replacement for complex semantics), with sufficient category diversity to support this focus. Full coverage of all semantic categories is left as future work.

Beyond $R@1$, we can include higher values of k in $R@k$, but this may introduce ambiguity in interpreting MA-CIR results, as candidate images differ only in localized regions. Namely, this can lead to visually similar but semantically incorrect retrievals being counted as correct, inflating scores and potentially misleading fine-grained evaluations of compositional understanding. For this reason, we do not report $R@k$ beyond $R@1$.

Although we employ a refinement and evaluation process to obtain high-quality images, there may still be inherent biases introduced by the image generation (or inpainting) model as well as the human-in-the-loop selection and refinement procedures. We leave more systematic bias analysis and mitigation strategies as future work.

While our simple remedy improves E5-V on MA-CIR in all experiments, the performance gain in Tab. 2 is relatively smaller than Tab. 1. This suggests that the remedy is particularly effective for improving robustness to hard negatives, whereas further gains in understanding complex semantics may require more carefully constructed text triplets or more advanced methods.

References

- [1] Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. In *ICLR*, 2024. 1, 4
- [2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *CVPR*, 2022. 4
- [3] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, 2023. 4
- [4] Jaeseok Byun, Seokhyeon Jeong, Wonjae Kim, Sanghyuk Chun, and Taesup Moon. Reducing task discrepancy of text encoders for zero-shot composed image retrieval. *arXiv preprint arXiv:2406.09188*, 2024. 4
- [5] Geonmo Gu, Sanghyuk Chun, HeeJae Jun, Yoohoon Kang, Wonjae Kim, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *TMLR*, 2023. 4
- [6] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, , Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. In *CVPR*, 2024. 1, 3, 4
- [7] Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. In *ECCV*, 2024. 1, 4
- [8] 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. 3, 4
- [9] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *ICLR*, 2024. 1, 4
- [10] Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan Kankanhalli. Improving context understanding in multimodal large language models via multimodal composition learning. In *ICML*, 2024. 1, 2
- [11] Haoqiang Lin, Haokun Wen, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. Fine-grained textual inversion network for zero-shot composed image retrieval. In *SIGIR*, 2024. 4
- [12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 3
- [13] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *WACV*, 2024. 4
- [14] Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. Preserving multi-modal capabilities of pre-trained vlms for improving vision-linguistic compositionality. In *EMNLP*, 2024. 1
- [15] Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *CVPR*, 2024. 4
- [16] 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 *CVPR*, 2023. 4
- [17] 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 *AAAI*, 2024. 4
- [18] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. CoVR: Learning composed video retrieval from web video captions. In *AAAI*, 2024. 1, 2
- [19] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2023. 1
- [20] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. In *ICML*, 2024. 1, 4