WISER: Wider Search, Deeper Thinking, and Adaptive Fusion for Training-Free Zero-Shot Composed Image Retrieval

Supplementary Material

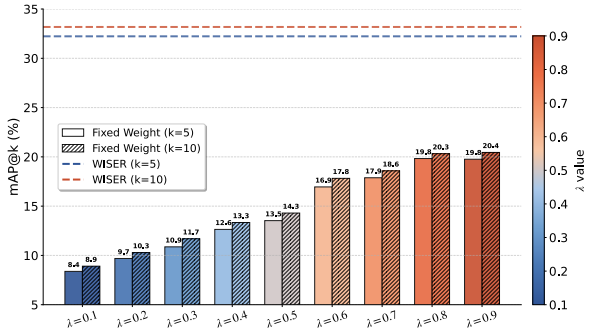

Figure 5. **Comparison between fixed fusion strategies and WISER on CIRCO.** λ controls the T2I weight in the fixed fusion (λ for T2I and 1-λ for I2I). Our WISER method achieves superior performance over all λ values, highlighting the limitation of static weighting.

6. More Ablation Study

To further validate each component's contribution, we conduct additional ablations on the reranking component of Adaptive Fusion using single-path baselines (Table 5). Applied to the T2I-only baseline, it improves R@10 by 7.78% on Fashion-IQ and mAP@5 by 11.12% on CIRCO; for the I2I-only baseline, the gains are 9.99% and 13.40%, respectively. This shows that the verifier effectively identifies target candidates within each pathway's retrieval results. However, a significant gap remains compared to the full WISER framework, indicating that while reranking helps, the complementary strengths of T2I and I2I are essential for optimal performance in diverse CIR scenarios.

We further test other refiners (e.g, DeepSeek-R1, Gemini 2.5 Pro) in Table 6. Both models achieve strong performance comparable to our default setup, confirming that WISER's gains stem from the framework design rather than reliance on a specific model. We empirically set the threshold to 0.7 and selected GPT-4o as the refiner without extra tuning. We believe targeted tuning could further improve performance, and thus report only the base results here.

7. More Qualitative Examples

In this section, we show more qualitative results on three datasets. We present the multimodal query, along with the edited image, edited caption, suggestions, and the top-5 retrieval results from T2I, I2I, and our method.

Table 5. **More ablation study on the core components of WISER on the Fashion-IQ and CIRCO datasets.** "RAK" denotes rerank, "AVG" denotes average fusion at the similarity level, and "ADA" denotes Adaptive Fusion.

Wider Search		Fusion	Deeper Thinking	Fashion-IQ-Avg		CIRCO			
T2I	I2I			R@10	R@50	mAP@5	mAP@10	mAP@25	mAP@50
-	✓	-	-	22.65	38.84	7.00	7.46	8.40	8.95
-	✓	-	✓	23.58	40.10	7.57	8.05	9.04	9.62
-	✓	RAK	-	32.64	38.84	20.40	20.12	20.63	20.73
✓	-	-	-	28.59	49.18	17.28	17.94	19.64	20.51
✓	-	-	✓	29.22	49.94	17.64	18.30	20.01	20.94
✓	-	RAK	-	36.37	49.18	28.40	28.74	30.22	30.56
✓	✓	AVG	-	33.40	52.92	13.53	14.30	15.96	16.77
✓	✓	ADA	-	40.83	57.86	31.32	32.08	33.72	34.24
✓	✓	ADA	✓	41.99	58.74	32.23	33.18	34.82	35.35

Table 6. **Further ablations of the refiner on CIRCO.**

Method	k=5	k=10	k=25	k=50
w/ Gemini 2.5 Pro	32.92	33.65	35.4	35.95
w/ DeepSeek-R1	33.08	33.85	35.51	36.05
WISER (Ours)	32.23	33.18	34.82	35.35

7.1. More Qualitative results on CIRR

We present additional qualitative results on CIRR in Figure 7. In Case 1, WISER combines the strengths of I2I to capture fine-grained visual details (e.g., the dog's breed) and T2I for semantic understanding, thereby retrieving the target image at top-1. In Case 2, although I2I generates a reasonable edited image that follows the modification intent, the inherent fuzziness of ZS-CIR (e.g., variations in the dog's orientation) introduces retrieval challenges. T2I retrieves more relevant candidates, benefiting from a more flexible textual representation. By adaptively fusing both pathways, WISER achieves top-1 retrieval of the target image. Case 3 represents a highly complex modification where the target image has a weak correlation with the reference image. Both T2I and I2I baselines are misled by visual information from the reference, leading to initial retrieval failure. Consequently, this triggers Deeper Thinking to refine the edited image. Although some noise remains, WISER demonstrates strong robustness by successfully identifying the target image at top-1, highlighting its ability to handle abstract and semantically challenging edits.

7.2. More Qualitative results on CIRCO

We present additional qualitative results on CIRCO in Figure 8. In Case 1, the modification intent is inherently am-

biguous. I2I fails to retrieve the target, likely due to its strict reliance on visual similarity. In contrast, T2I successfully retrieves two target images within the top-5 by capturing key semantic elements while allowing for visual variation. WISER further expands retrieval diversity and returns more relevant targets through its adaptive fusion. Case 2 demonstrates a scenario where visual precision is critical. I2I excels by preserving fine-grained details (e.g., the bird's breed) and successfully retrieves the target image at top-1. T2I suffers from the inherent ambiguity of textual representation and fails to identify the correct instance. WISER maintains the strong performance of I2I, highlighting its ability to preserve visual fidelity when it is essential. Case 3 involves a complex compositional edit. Initially, both I2I and T2I struggle: I2I fails to generate a correct edited image ("two people on the same bike"), while T2I does not fully capture the precise semantic constraint ("on the same bike"). This uncertainty triggers Deeper Thinking. After refinement, both the edited image and caption accurately reflect the intended modification, enabling WISER to retrieve the target image at top-1 correctly. This case highlights the critical role of iterative refinement in resolving semantically and visually challenging queries.

7.3. More Qualitative results on Fashion-IQ

We present additional qualitative results on the Fashion-IQ dataset in Figure 9. In Case 1, due to the ambiguity in translating the specific attributes into a visual edit, I2I fails to retrieve the target accurately. In contrast, both T2I and WISER successfully retrieve the target image at top-1, demonstrating the advantage of semantic understanding in capturing detailed attribute-based changes. In Case 2, I2I excels by preserving the structural details of the reference garment while accurately applying the color and pattern modifications, leading to correct top-1 retrieval. WISER maintains this strong performance through adaptive fusion. Case 3 presents a more complex color transformation. I2I fails to generate a correct color gradient, while T2I introduces interference from the reference image by retaining the "red and blue plaid" pattern in its edited caption. This imprecision leads to retrieval inaccuracy. WISER, however, identifies the uncertainty and triggers Deeper Thinking to improve retrieval performance.

7.4. Failure Cases

We also demonstrate failure cases across three datasets in Figure 10. For Case 1 from CIRCO, I2I retains most visual information from the reference image and correctly generates the edited scene, which helps retrieve the target within top-5. However, it also introduces distraction by prioritizing stylistically similar but semantically unmatched images. T2I misunderstands the reference scene, incorrectly describing "a man is working on a laptop" instead of "a DJ facing the camera with a console and a laptop." WISER overemphasizes the "microphone" attribute while neglecting other contextual information, leading to retrieval inaccuracy. In Case 2 from CIRR, the modification requires complete replacement of the main subject, presenting a significant challenge. I2I mistakenly retains the entity count, generating two dogs instead of one. Although T2I captures the modification intent correctly, and WISER successfully refines the edited image after one iteration, all methods ultimately fail due to the high noise and inherent ambiguity in the CIRR dataset. Case 3 from Fashion-IQ involves an ambiguous modification request. Due to the subjective nature of the description and the high visual similarity among fashion items, both T2I and WISER struggle to precisely identify the target from a large pool of candidate images with similar attributes. This case underscores the difficulty in handling subjective or abstract attribute changes within a fine-grained retrieval domain.

Prompt for the verifier

You are a strict visual verifier. Output exactly one token: yes or no (lowercase). Do not add punctuation or explanations.
Reference image: <reference image>
Candidate image: <candidate image>
Instruction:{modification text}

Decide if the candidate image matches the result of applying the instruction to the reference image.

Return yes if all required elements implied by the instruction are satisfied (like counts, categories, attributes, spatial relations). If any required element is missing or contradicted, answer no.

Answer:"""

Figure 6. **Prompt for the verifier.** The prompt guides the verifier to answer a binary question given the reference image, the candidate image and the modification text.

8. Prompt

In this section, we illustrate all the prompts used in our paper. For Adaptive Fusion, we use the prompt shown in Figure 6. For Deeper Thinking, the prompt for T2I and I2I is shown in Figure 11 and Figure 12, respectively. Part of our prompts are taken from AutoCIR [22].

Figure 7. **More Qualitative Results on CIRR.** Green checkmark indicates correct target retrieval. WISER successfully retrieves the target image at top-1 across various scenarios.

Figure 8. **More Qualitative Results on CIRCO.** Green checkmark indicates correct target retrieval. WISER successfully retrieves the target image at top-1 across various scenarios.

Figure 9. **More Qualitative Results on Fashion-IQ.** Green checkmark indicates correct target retrieval. WISER successfully retrieves the target image at top-1 across various scenarios.

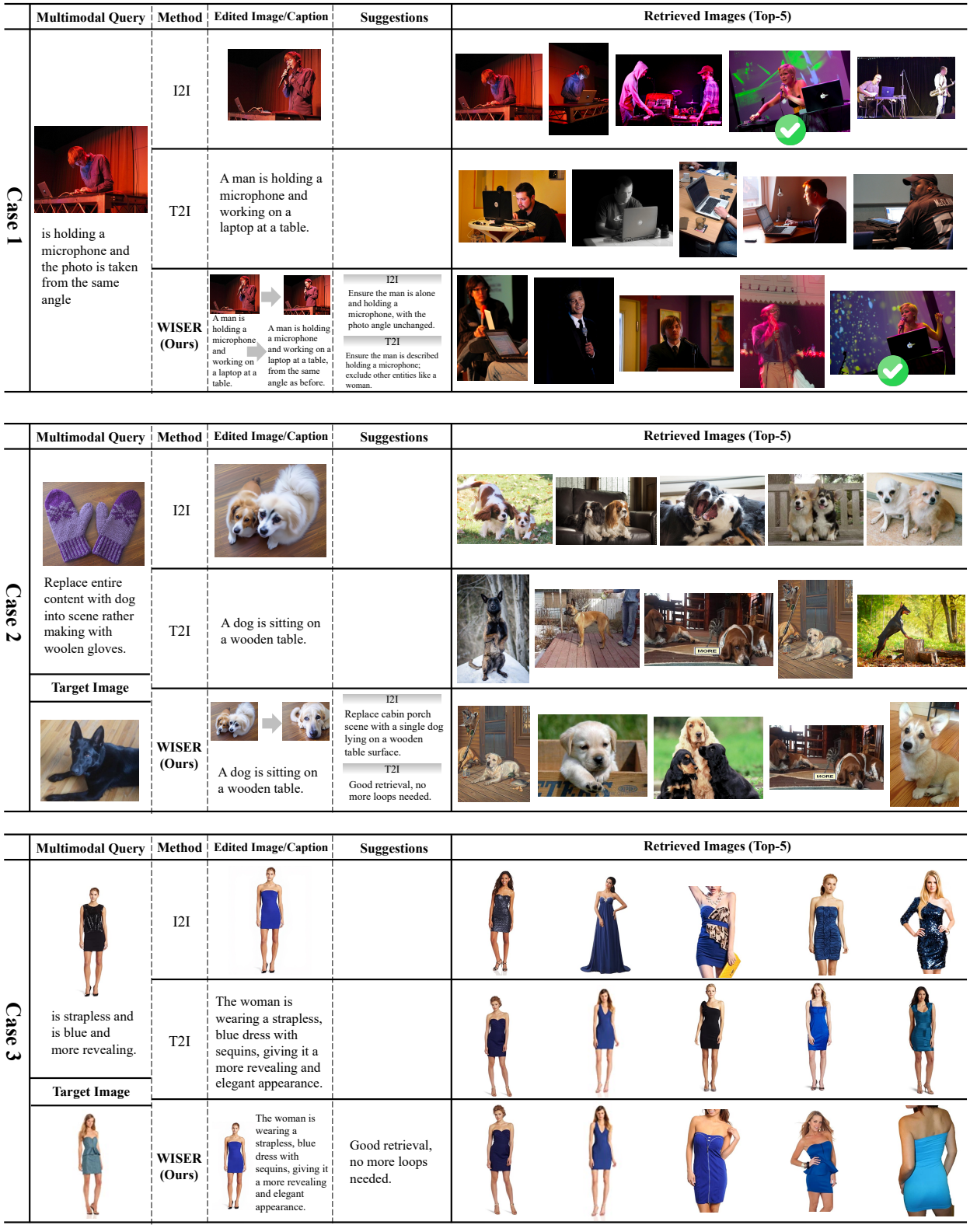

Figure 10. **Failure cases from CIRCO, CIRR, Fashion-IQ datasets.** We demonstrate three challenging cases where WISER struggles to retrieve the target image at top-1.

Prompt for the refiner (T2I)

Assume you are an experienced composed image retrieval expert, skilled at precisely generating new image descriptions based on a reference image's description and the user's modification instructions.
You excel at creating modified descriptions that can retrieve images matching the user's requested changes through vector retrieval.
Your task is to help improve the effectiveness of compositional image retrieval by generating precise modification suggestions that will assist another large language model (LLM) in producing a better image description.
Please note that this LLM has received the reference image's description and the user's modification instructions, and already generated a modified description.
Moreover, a retrieval has been performed based on this modified description. Thus your task is to analyze the last retrieval result and provide modification suggestions and please follow the below steps to finish this task.

Step 1: Identifying Modifications
Your first task is to identify the modifications and generate corresponding modification phrases.
Specifically, here is the description of the reference image: "{reference image caption}." Here are the user's modification requests: "{modification text}"
By deeply understanding the image description and the user's modifications, please generate the following two types of modification phrases:
1. If the modification involves changing the characteristics of an entity in the original reference image, please specify the changes,
2. If the modification involves adding or deleting an entity, please specify the additions or deletions.
Please note that the user's modifications may lack a subject; in such cases, infer and supply the object corresponding to the modification.
Only include modifications explicitly mentioned by the user. If a certain type of modification is not present, you do not need to provide it and should avoid generating unspecified content.

Step 2: Analyzing the Retrieved Image
Compare the modification phrases identified in Step 1 with the description of the retrieved image : "{pseudo image caption}". Note that this retrieval is performed with the modified description generated by another LLM, which has been mentioned above.
Determine if the retrieved image meets the user's modification instructions.
If it matches after excluding subjective modifications (e.g., "casual," "relaxed"), respond with: "Good retrieval, no more loops needed."
If there are unmet modification phrases, proceed to Step 3.

Step 3: Providing Modification Suggestions
For any unmet modifications identified in Step 2, suggest targeted changes to help the LLM regenerate an improved modified description. Keep suggestions concise and specific to ensure they effectively guide the LLM.
Output format:
"Suggestion: <concise, actionable suggestion in 10-20 words>"

Figure 11. **Prompt for the refiner of T2I.** The prompt guides the refiner to perform structured self-reflection for uncertain retrievals, given the reference image caption, modification text, and pseudo target caption from T2I. The generated suggestions are then fed to the editor to refine the edited caption for T2I.

Prompt for the refiner (I2I)

Assume you are an experienced composed image retrieval expert, skilled at precisely generating new image based on a reference image's description and the user's modification instructions.
You excel at creating modified images that can retrieve images matching the user's requested changes through vector retrieval.
Your task is to help improve the effectiveness of compositional image retrieval by generating precise modification suggestions that will assist another multimodal large language model (MLLM) in producing a better image.
Please note that this MLLM has received the reference image's description and the user's modification instructions, and already generated a modified image.
Moreover, a retrieval has been performed based on this modified image. Thus your task is to analyze the last retrieval result and provide modification suggestions and please follow the below steps to finish this task.

Step 1: Identifying Modifications
Your first task is to identify the modifications and generate corresponding modification phrases.
Specifically, here is the description of the reference image: "{reference image caption}." Here are the user's modification requests: "{modification text}."
By deeply understanding the image description and the user's modifications, please generate the following two types of modification phrases:
1. If the modification involves changing the characteristics of an entity in the original reference image, please specify the changes,
2. If the modification involves adding or deleting an entity, please specify the additions or deletions.
Please note that the user's modifications may lack a subject; in such cases, infer and supply the object corresponding to the modification.
Only include modifications explicitly mentioned by the user. If a certain type of modification is not present, you do not need to provide it and should avoid generating unspecified content.

Step 2: Analyzing the Retrieved Image
Compare the modification phrases identified in Step 1 with the description of the retrieved image : "{}". Note that this retrieval is performed with the modified image generated by another MLLM, which has been mentioned above.
Determine if the retrieved image meets the user's modification instructions.
If it matches after excluding subjective modifications (e.g., "casual," "relaxed"), respond with: "Good retrieval, no more loops needed."
If there are unmet modification phrases, proceed to Step 3.

Step 3: Providing Modification Suggestions
For any unmet modifications identified in Step 2, suggest targeted changes to help the MLLM regenerate an improved modified image. Keep suggestions concise and specific to ensure they effectively guide the MLLM.
Output format:
"Suggestion: <concise, actionable suggestion in 10-20 words>"

Figure 12. **Prompt for the refiner of I2I.** The prompt guides the refiner to perform structured self-reflection for uncertain retrievals, given the reference image caption, modification text, and pseudo target caption from I2I. The generated suggestions are then fed to the editor to refine the edited image for I2I.