Beyond Semantic Search: Towards Referential Anchoring in Composed Image Retrieval

Supplementary Material

1. More Details on the OACIRR Benchmark

In this section, we provide a comprehensive overview of the construction pipeline and detailed statistics of the **OACIRR** benchmark. We describe the subset-specific protocols in Section 1.1, the prompts used for MLLM-based annotation in Section 1.2, the detailed dataset statistics in Section 1.3, and the instance diversity visualization in Section 1.4.

1.1. Subset-Specific Construction Pipeline

We construct the four **OACIRR** subsets — Fashion, Car, Product, and Landmark — using four large-scale, fine-grained visual classification datasets: DeepFashion2 [5], Stanford Cars [8], Products-10K [2], and Google Landmarks v2 [17]. Given that these sources differ substantially in structure and granularity, we design tailored protocols and apply subset-specific filtering thresholds throughout the construction pipeline. We detail each stage below:

Stage 1: Image Pair Collection. The objective of this stage is to establish high-fidelity, instance-level image sets, with procedures tailored to each data source:

- For **Products-10K**, the images are already organized at the stock-keeping-unit (SKU) level, which naturally aligns with our instance-level fidelity requirement.
- For **DeepFashion2** and **Stanford Cars**, the initial groupings (based on item styles or car models) often contain multiple color variants. To obtain color-consistent instance sets, we further subdivide each group using a pre-trained fine-grained classifier (CLIP-ConvNeXt-Base).
- For **Google Landmarks v2**, image sets vary between *visually coherent views* of a landmark and *knowledge-based collections* that mix disparate appearances. To enforce strict visual consistency, we prompt an MLLM [1] to identify and retain only visually coherent subsets.

Stage 2: Image Pair Filtering. As summarized in Table 1, we apply subset-specific thresholds to ensure high-quality image pairs and appropriate task difficulty. A set $\mathcal{S}_j$ is retained only if its size exceeds the construction-valid threshold τ_{valid}. Image pairs with feature cosine similarity above τ_{high} are removed to ensure meaningful modifications. To promote background diversity, an image is filtered out if its feature similarity exceeds $\tau_{centric}$ with at least τ_{count} other images in the same set.

- To balance the query volume across domains, we adopt smaller τ_{valid} values for subsets with fewer initial IDs (Fashion, Car) and larger values for subsets with abundant initial IDs (Product, Landmark).

Subset	Filtering Threshold			
	τ_{valid}	τ_{high}	$\tau_{centric}$	τ_{count}
Fashion	8	0.92	0.88	3
Car	10	0.88	0.85	2
Product	20	0.88	0.85	2
Landmark	15	0.90	0.88	3

Table 1. Filtering Thresholds for each **OACIRR** subset.

- To calibrate task difficulty across domains, we adopt more relaxed thresholds (τ_{high}, $\tau_{centric}$, τ_{count}) for subsets involving complex multi-object scenes (Fashion, Landmark), and more rigorous thresholds for subsets centered around a single salient object (Car, Product).

Stage 3: Quadruple Annotation. This stage involves a semi-automatic process. We assign class labels l_{ins} to each high-fidelity instance set using a tailored prompt. To reinforce the synergy between the visual and textual modalities, we instruct the MLLM to generate modification texts describing only contextual changes, explicitly excluding any mention of the preserved instance. For bounding boxes, we directly use the ground-truth annotations in DeepFashion2. For the remaining three subsets, bounding box proposals with confidence scores below 0.3 from our grounding model [20] are manually re-annotated to ensure precision.

Stage 4: Candidate Gallery Construction. To construct challenging yet efficient candidate galleries, we compute the instance class distribution for each test subset. Each gallery is populated by sampling hard negatives from the reserved image pool (from Stage 1) to match the class distribution of the query set. This strategy maximizes instance-level ambiguity while maintaining a compact and computationally efficient gallery for the benchmark.

1.2. MLLM Annotation Prompts

We employed Qwen-VL-Max [1] for all MLLM-based annotation tasks, which comprise two key sub-tasks: (1) *generating class labels* for each high-fidelity instance set, and (2) *producing contextual modification text* conditioned on an image pair and its associated instance class label.

Instance Class Label Generation. This step was applied selectively depending on the characteristics of each subset. For the Fashion subset, we directly adopted the coarse-grained apparel categories defined in DeepFashion2. For the Car subset, all instances were uniformly assigned the label "`car`". Consequently, MLLM-based labeling was re-

quired only for the Product and Landmark subsets, which exhibit greater category diversity.

For the Product subset, which involves only class label annotation, the following prompt template was used:

Class Label Generation for Product subset

Analyze the provided images to identify the single, identical commercial product present in all of them. Your task is to output a concise, generic tag for this common object.

Important Context:
1. There is exactly one object that is the same product across all images.
2. This object may appear in different states, environments, or from different viewing angles in each image.

Requirements:
1. Output only the tag for the common object and nothing else.
2. The tag must be a short, descriptive noun phrase in English. It should be specific enough to be unambiguous but not overly detailed.
3. DO NOT include any brand names.
4. DO NOT describe the object's state, its background, the viewing angle, or any similarities or differences between the images.
5. DO NOT include any introductory phrases like "The common object is:".

For the Landmark subset, we designed a prompt that concurrently performs visual consistency filtering and class label annotation. The prompt template is as follows:

Visual Consistency Filter & Class Label Generation for Landmark subset

Your task is to analyze a set of images from a single landmark ID and determine if they represent a "Visual-type" or a "Knowledge-type" landmark, based ONLY on the visual evidence provided.
When in doubt, classify as "Knowledge-type". Your goal is to approve "Visual-type" only when the images unambiguously represent a single, consistent landmark, with verification purely from visual cues.

Landmark Types Explained:
1. *Visual-type*: The images depict a single, visually consistent, and dominant landmark. The landmark is the same physical entity across all images, even when viewed from different angles or under varying conditions (e.g., day/night, summer/winter).
2. *Knowledge-type*: The images are related by a shared theme or geographic context but do not contain one visually consistent landmark. Their connection is conceptual or requires external knowledge to identify. (e.g., different buildings within a university campus; interior and exterior views of a large museum.)

Response Format:
Your response MUST be a JSON object and nothing else. Follow this exact format:
{
"type": "visual" or "knowledge",
"label": " Specific Name of the Landmark" or null,
"reasoning": "A brief explanation for your decision."
}

Important Rules:
1. If you classify as "knowledge", set "label" to null.
2. If you classify as "visual", provide the class label of the landmark for the "label".
3. Do not include any introductory text before or after the JSON object.

Contextual Modification Text Generation. To ensure that the generated modification text is accurate, diverse, and effectively complements the visual information, we designed domain-specific prompt templates for all four subsets. A shared instruction across these prompts was to restrict the MLLM to describe only contextual changes, thereby maximizing its synergy with the visual anchor. The corresponding prompt templates are provided below.

Modification Text Generation for Fashion subset

Based on the two provided images, generate a modification text to transform the first image into the second.

Requirements:
1. The modification text must be written in fluent and natural English, NOT exceeding 30 words.
2. Focus exclusively on the most significant and definite changes. DO NOT describe any identical parts between the two images.
3. A specific " Object to Ignore" is provided below. DO NOT mention this object or any of its attributes in the modification text.
4. Avoid any explicit references to the images themselves. For example, DO NOT use phrases like " in the first image" or " in the second picture".
5. Employ diverse expressions. Avoid using repetitive sentence structures or fixed grammatical patterns.

Examples:
1. The woman is now wearing a large pink bow and holding a light-up wand.
2. The person is wearing a denim skirt, and the background changes to a store with shelves and products.
3. The girl changed from wearing patterned pants to white cut-off shorts, and moved from an indoor yoga room to an outdoor pathway.

Object to Ignore: `[Object]`

Modification Text Generation for Car subset

Based on the two provided images, generate a modification text that describes the changes from the first image to the second.

Important Context:
The car (model and color) is the same in both images.

Requirements:
1. The modification text must be written in fluent and natural English, NOT exceeding 25 words.
2. Focus exclusively on the most significant and definite changes (e.g., Background / Environment, Viewing Angle, Car's State). DO NOT describe the car's model or color, as they are unchanged.
3. Avoid any explicit references to the images themselves. For example, DO NOT use phrases like "in the first image" or "in the second picture".
4. Employ diverse expressions. Avoid using repetitive sentence structures or fixed grammatical patterns.

Examples:
1. Now shown from a low-angle perspective.
2. The scene changes to a desert at sunset.
3. The car is now viewed from a front angle on a snowy mountain road with its headlights turned on.
4. Instead of being parked in a garage, the vehicle is now on a bridge with its driver-side door open.

Modification Text Generation for Product subset

Based on the two provided images, generate a modification text that describes the changes from the first image to the second.

Important Context:
The product object: `[Object]` is the same in both images. You are strictly forbidden from mentioning this product in your response. Your task is to describe how its presentation has changed.

Requirements:
1. The modification text must be written in fluent and natural English, NOT exceeding 30 words.
2. Focus exclusively on the most significant and definite changes (e.g., Background / Environment, Viewing Angle, State, Packaging, Interaction).
3. A specific "Object to Ignore" is provided below. DO NOT mention this product object or any of its attributes (e.g., color, brand, type) in your response.
4. Avoid any explicit references to the images themselves. For example, DO NOT use phrases like "in the first image" or "in the second picture".
5. Employ diverse expressions. Avoid using repetitive sentence structures or fixed grammatical patterns.

Examples:
1. Now shown from a top-down perspective.
2. Now shown out of its original packaging.
3. The laptop is open and displayed on a wooden desk.
4. The sneakers are now being worn by a person on a basketball court.

Object to Ignore: `[Object]`

ModificationText Generation for Landmark subset

Based on the two provided images, generate a modification text to transform the first image into the second.

Important Context:
Both images are about the landmark: `[Object]`. You are strictly forbidden from mentioning this landmark in your response. Your task is to describe how its context, framing, and atmosphere has changed.

Requirements:
1. The modification text must be written in fluent and natural English, NOT exceeding 30 words.
2. Focus exclusively on the most significant and definite changes (e.g., Viewing Angle, Change in Scope or Focus, Atmospheric Conditions, Surrounding Environment).
3. A specific "Object to Ignore" is provided below. DO NOT mention this landmark, its name, its architectural style, or its location in your response.
4. Avoid any explicit references to the images themselves. For example, DO NOT use phrases like "in the first image" or "in the second picture".

5. Employ diverse expressions. Avoid using repetitive sentence structures or fixed grammatical patterns.

Examples:
1. Now seen from an aerial perspective on a clear day.
2. The scene shifts to a clear night, with the structure illuminated.
3. Now viewed from across the river on a foggy morning, with autumn foliage visible.

Object to Ignore: `[Object]`

Statistic	Number	Percentage
Total Annotated Quadruples	**127,166**	
- Fashion	12,874	10.1%
- Car	12,728	10.0%
- Product	75,616	59.5%
- Landmark	25,948	20.4%
Total Unique Images	**39,495**	
- Fashion	1,034	2.6%
- Car	3,111	7.9%
- Product	27,531	69.7%
- Landmark	7,819	19.8%
Total Unique Instances	**2,647**	
- Fashion	80	3.0%
- Car	199	7.5%
- Product	1,419	53.6%
- Landmark	949	35.9%
Maximum Modification Text Length	30.0	-
Average Modification Text Length	20.2	-

Table 2. Statistics of **OACIRR** Training Dataset.

Statistic	Number	Percentage
Total Annotated Quadruples	**33,449**	
- Fashion	3,606	10.8%
- Car	3,586	10.7%
- Product	21,046	62.9%
- Landmark	5,211	15.6%
Total Unique Images	**26,595**	
Quadruple Images	15,467	58.1%
Distractor Images	11,134	41.9%
- Fashion	5,077	19.1%
- Car	4,717	17.7%
- Product	11,801	44.4%
- Landmark	5,000	18.8%
Total Unique Instances	**4,945**	
Quadruple Instances	1,238	25.0%
Distractor Instances	3,707	75.0%
- Fashion	1,683	34.0%
- Car	1,089	22.0%
- Product	799	16.2%
- Landmark	1,374	27.8%
Maximum Modification Text Length	30.0	-
Average Modification Text Length	19.4	-

Table 3. Statistics of **OACIRR** Evaluation Benchmark.

1.3. Detailed Dataset Statistics

As shown in Tables 2 and 3, we provide a detailed statistical breakdown of the **OACIRR** benchmark, highlighting the scale and diversity of both the training data and the evaluation benchmark. The partitioning and design of **OACIRR** were guided by two principles to ensure rigor and utility:

- **Strict data partitioning for fair evaluation.** We enforce a strict separation between the training and evaluation splits by ensuring that no images or instances overlap between them. We further reduce fine-grained category overlap to prevent data leakage and ensure that evaluation faithfully reflects generalization to unseen instances.
- **Asymmetric design for comprehensive evaluation.** The asymmetric composition of the four subsets is a deliberate design choice that leverages domain-specific characteristics to assess complementary retrieval capabilities. The Fashion, Car, and Landmark subsets emphasize *retrieval depth*, requiring discrimination among visually similar instances within a coherent domain. In contrast, the Product subset targets *retrieval breadth*, evaluating robustness under substantially larger and more diverse candidate spaces. Collectively, these complementary settings provide a holistic assessment of both fine-grained discrimination and large-scale retrieval performance.

1.4. Instance Diversity Visualization

Figure 1 presents a curated collage of representative, cropped instances from the four primary domains, offering a compact visual summary of the benchmark's scope. **OACIRR** covers a broad spectrum of categories, ranging from everyday apparel and common vehicles to diverse consumer goods and iconic global sites, exposing models to a wide variety of visual concepts and real-world contexts.

Complementing this breadth, **OACIRR** also exhibits substantial fine-grained depth. Individual sub-categories are densely populated with numerous distinct instances, encompassing a wide range of appearance variations. Such granularity enables evaluation to extend beyond coarse category recognition toward precise, instance-level discrimination. Collectively, this diversity and depth establish **OACIRR** as a comprehensive and challenging benchmark for instance-aware compositional retrieval.

2. Additional Evaluation Protocols and Results

To supplement the quantitative results in the main text, this section provides the detailed evaluation protocols used to adapt existing retrieval paradigms to the OACIR task and presents additional results under alternative configurations. Section 2.1 details the two adaptation settings that convert the anchored-instance constraint into formats compatible with different model architectures, and Section 2.2 reports supplementary quantitative results under these settings.

2.1. Details on Evaluation Protocols

Setting 1: Instance-as-Textual Adaptation. The anchored object is specified through a textual cue. A short template containing the instance's class label is appended to the original modification text, converting the OACIR task into an instance-aware CIR formulation while preserving richer contextual information. This setting assesses the model's capacity to ground fine-grained textual constraints within a visually complex query. Prompt templates are given below:

Prompt Templates for Setting 1

1. Same `[Object]`
2. With the same `[Object]`
3. Fixed `[Object]`
4. Identical `[Object]`
5. Invariant `[Object]`
6. Keep the `[Object]`
7. Preserving the `[Object]`
8. `[Object]` unchanged

Setting 2: Instance-as-Visual Adaptation. The anchored object is provided as an explicit visual cue by rendering its bounding box onto the reference image and pairing it with a brief instruction. This setting assesses the model's capacity to interpret direct visual grounding signals for instance preservation. The instruction is given below:

Instruction for Setting 2

`[Prompt Template for Setting 1]`
in the `[Color]` bounding box.

Model-Specific Application. Universal Multimodal Retrieval (UMR) models rely heavily on visual grounding and instructional prompts. Therefore, we adopt **Setting 2** as the default protocol for these models, using domain–specific instructions tailored to each OACIRR subset. The complete domain-specific instruction templates are provided below.

UMR Instruction Templates for Fashion subset

1. Find a fashion image that aligns with the reference image and style note.

2. Retrieve a fashion scene image that reflects the described transformation from the provided image.

3. Can you find an outfit image that meets the adjustments described in the text?

4. I'm looking for a similar fashion image with the described changes to the model and scene.

UMR Instruction Templates for Car subset

1. Retrieve a car image that aligns with the reference image and the scene modifications.

2. Find a vehicle image like this one, but with the adjustments from the text.

3. Can you pull up a car image that incorporates the requested changes?

4. I'm looking for a similar car image with the described changes to the setting and angle.

UMR Instruction Templates for Product subset

1. Find a product image that aligns with the provided image and the modification instructions.

2. Given the reference image and display notes, find the matching product image.

3. Can you find a product image that meets the requested changes to the background and view?

4. I'm looking for a similar product image matches the new display style from the text.

UMR Instruction Templates for Landmark subset

1. Retrieve a landmark image that aligns with the reference image and the described conditions.

2. Pull up a photo of a landmark that matches the reference image and the requested transformation.

3. Given the reference image and description, identify the corresponding landmark view.

4. I'm looking for a similar landmark image with the specified changes in atmosphere and perspective.

In contrast, Zero-shot and Supervised CIR methods do not support bounding-box inputs. Therefore, we adopt **Setting 1** as their default protocol, translating the instance constraint into a textual form compatible with their workflow.

2.2. Ablation on Evaluation Protocols

To validate these choices, we additionally evaluate UMR models under *Setting 1* and CIR models under *Setting 2*. As shown in Table 4, each model class performs best under its default protocol, indicating that UMR models rely on explicit visual grounding while CIR models favor semantically integrated textual cues. In contrast, our ***AdaFocal*** provides a robust encoding mechanism that adapts reliably to the OACIR task and its anchored-instance constraint.

Domain	Method	Pretraining Data	Fashion			Car			Product			Landmark			*Avg.*
			R_{ID}@1	R@1	R@5	R_{ID}@1	R@1	R@5	R_{ID}@1	R@1	R@5	R_{ID}@1	R@1	R@5	
Setting 1: Instance-as-Textual Adaptation															
UMR	LamRA-Ret [13]	M-BEIR + NLI	25.93	20.54	36.26	58.13	33.87	72.10	67.27	36.64	67.51	57.05	32.06	67.99	47.95
	MM-Embed [12]	M-BEIR + MTEB	38.05	32.70	50.69	51.37	29.62	61.74	66.68	36.73	65.49	75.95	37.75	78.53	52.11
	GME (2B) [19]	UMRB	37.10	31.45	51.33	55.91	30.37	63.94	75.91	40.90	72.39	72.65	38.76	74.46	53.76
	GME (7B) [19]		44.54	38.33	59.51	58.73	35.05	70.91	81.87	53.42	82.97	76.20	46.82	82.27	60.89
	U-MARVEL [11]	M-BEIR + NLI	44.32	39.14	59.64	59.63	38.17	72.16	80.78	51.40	81.01	68.00	37.08	72.23	58.63
ZS-CIR	Pic2Word [15]	CC3M	14.98	11.15	21.55	12.07	4.07	11.32	45.95	13.66	34.19	55.98	20.99	52.12	24.84
	LinCIR [7]		15.78	12.04	21.82	5.55	2.23	7.28	47.55	14.63	34.91	42.76	19.57	47.15	22.61
CIR	SPRC (ViT-G) [3]	CIRR	28.62	25.79	44.48	25.13	15.92	37.06	54.39	34.85	62.31	40.41	26.29	52.39	37.30
		OACIRR (Ours)	65.25	58.51	80.89	72.87	49.82	89.57	86.05	70.61	93.68	76.32	56.04	89.00	74.05
Setting 2: Instance-as-Visual Adaptation															
UMR	LamRA-Ret [13]	M-BEIR + NLI	27.45	21.63	37.10	61.03	35.44	74.51	69.45	39.53	70.25	58.64	32.58	68.74	49.70
	MM-Embed [12]	M-BEIR + MTEB	41.38	34.55	52.50	53.21	30.06	62.80	71.03	41.47	71.15	78.85	38.88	79.32	54.60
	GME (2B) [19]	UMRB	38.13	32.14	51.50	58.84	31.60	66.03	76.89	44.11	74.20	73.86	38.99	75.61	55.16
	GME (7B) [19]		44.98	39.24	60.18	63.11	38.34	75.38	83.44	54.60	84.15	77.11	47.09	82.69	62.53
	U-MARVEL [11]	M-BEIR + NLI	46.05	40.38	60.59	62.92	39.96	74.90	83.26	54.69	84.13	69.81	37.67	73.08	60.62
ZS-CIR	Pic2Word [15]	CC3M	14.96	11.00	21.16	12.04	3.95	11.07	39.39	11.50	27.13	46.60	18.39	46.49	21.97
	LinCIR [7]		15.76	11.99	21.48	5.54	2.17	7.25	46.57	13.85	33.96	42.16	19.09	47.11	22.24
CIR	SPRC (ViT-G) [3]	CIRR	28.59	25.68	43.68	24.23	15.48	36.25	46.62	29.33	49.44	33.64	23.03	46.73	33.56
		OACIRR (Ours)	64.14	57.71	79.65	72.70	48.29	89.18	84.27	66.86	91.13	75.24	54.65	88.93	72.73
OACIR Task-Specific Architecture															
OACIR	**Baseline (ViT-G)**	**OACIRR (Ours)**	69.07	58.76	81.44	74.59	49.78	89.46	87.48	69.53	93.66	79.80	55.49	89.87	74.91
	Baseline† (ViT-G)		72.66	63.31	83.97	76.85	50.24	89.87	88.68	72.13	94.09	80.05	55.69	90.14	76.47
	SPRC† (ViT-G)		69.94	60.98	82.72	74.08	51.62	89.79	86.42	70.90	93.74	77.41	55.90	89.02	75.21
	***AdaFocal* (ViT-G)**		**77.15**	**65.31**	**86.88**	**78.42**	**53.63**	**92.22**	**91.86**	**74.11**	**95.39**	**82.92**	**58.47**	**91.63**	**79.00**

Table 4. **Quantitative comparison on the OACIRR benchmark under different evaluation settings and OACIR-specific baselines.** "Avg." represents the average results across all evaluation metrics. The best result is highlighted in **bold**, and the second best is underlined. **Baseline** denotes the standard CIR baseline, **Baseline**† denotes the ROI-cropped baseline, and **SPRC**† denotes plug-and-play CIR baseline.

3. Additional Ablation Studies

This section provides additional ablation studies to further validate the effectiveness of our method and the value of the **OACIRR** benchmark. Section 3.1 compares ***AdaFocal*** against stronger region-aware baselines, including explicit ROI cropping and a plug-and-play integration into SPRC [3], further verifying the effectiveness of our adaptive attention design. Section 3.2 and Section 3.3 evaluate the generalization ability of models trained on **OACIRR** across tasks and domains, respectively. Section 3.4 analyzes the robustness of ***AdaFocal*** to imperfect bounding box inputs. Finally, Section 3.5 examines key design choices within the *Context-Aware Attention Modulator (CAAM)*, including the modulation output form, the number of self-attention layers, and the configuration of contextual probe tokens.

3.1. Region-Aware Baselines

To rigorously evaluate the necessity and effectiveness of our adaptive attention mechanism, we compare ***AdaFocal*** against three region-aware baseline models, each reflecting a distinct way of handling the anchored instance constraint:

- **Standard CIR Baseline.** This model removes the *CAAM* module ($\beta = 0$) and encodes the full reference image and modification text using the Multimodal Encoder. While it preserves global visual context, it lacks any mechanism to preferentially attend to the anchored instance region.
- **ROI-Cropped Baseline (Baseline†).** To introduce explicit region awareness without additional learning, we crop the reference image using the bounding box B_r and feed only the cropped region into the encoder. This forces attention onto the instance but eliminates surrounding context essential for interpreting the modification text.
- **Plug-and-Play CIR Baseline (SPRC†).** To enable a fairer independent evaluation, we integrate the *CAAM* into the strong CIR model SPRC [3] by applying its dynamic attention activation during the first image-text fusion stage of the query encoder. This isolates *CAAM* as a plug-and-play module for instance-focused attention modulation within an existing CIR architecture.

As shown in Table 4, using the cropped instance (Baseline†) improves Instance Recall over the Standard Baseline, indicating that explicit isolation strengthens iden-

Method	Pretraining Data	Pretraining Scale	FashionIQ		CIRR				CIRCO		
			Avg@10	Avg@50	R@1	R@5	R_s@1	Avg.	mAP@5	mAP@10	mAP@25
CASE [9]	LaSCo + CoCo	389 K	–	–	35.40	65.78	64.29	65.04	–	–	–
CoVR-BLIP [16]	WebVid-CoVR	1,644 K	27.70	44.63	38.48	66.70	69.28	67.99	21.43	22.33	24.47
CompoDiff [6]	ST18M + LAION-2B	18,000 K	39.02	51.71	26.71	55.14	64.54	59.84	15.33	17.71	19.45
CoAlign (ViT-G) [10]	CIRHS	535 K	39.22	60.08	41.08	71.11	70.80	70.96	21.60	23.38	25.98
Baseline (ViT-L)	**OACIRR (Ours)**	**127 K**	37.82	59.57	41.59	72.36	72.06	72.21	21.51	23.14	25.47
Baseline (ViT-G)			**39.80**	**61.81**	**42.96**	**73.62**	**72.95**	**73.28**	**23.96**	**24.69**	**26.58**

Table 5. Zero-shot cross-task generalization on standard CIR benchmarks.

Setting	Method	Fashion			Car			Product			Landmark			*Avg.*
		R_{ID}@1	R@1	R@5	R_{ID}@1	R@1	R@5	R_{ID}@1	R@1	R@5	R_{ID}@1	R@1	R@5	
Cross-Domain	SPRC [3]	48.73	40.51	62.84	66.86	41.49	78.98	62.79	42.68	71.72	59.49	38.55	71.23	57.16
	AdaFocal	**61.15**	**50.25**	**71.54**	**74.26**	**45.65**	**85.81**	**67.84**	**45.53**	**74.68**	**61.58**	**40.66**	**71.94**	**62.57**
Full Finetuning	SPRC [3]	65.25	58.51	80.89	72.87	49.82	89.57	86.05	70.61	93.68	76.32	56.04	89.00	74.05
	AdaFocal	**77.15**	**65.31**	**86.88**	**78.42**	**53.63**	**92.22**	**91.86**	**74.11**	**95.39**	**82.92**	**58.47**	**91.63**	**79.00**

Table 6. Cross-domain generalization on the **OACIRR** benchmark.

tity preservation. However, its gains in standard Recall remain limited, suggesting that removing background context hinders the interpretation of contextual modifications. Integrating *CAAM* into SPRC (SPRC†) still improves over vanilla SPRC, verifying that our module is effective as a plug-and-play instance-aware attention mechanism. However, its gains remain limited, indicating that complex post-interaction layers prior to query encoding in existing CIR methods can dilute this direct instance-focused attention. In contrast, ***AdaFocal*** achieves the strongest overall balance between instance fidelity and compositional reasoning.

3.2. Cross-Task Generalization of OACIRR

We evaluate whether the instance-consistent supervision provided by **OACIRR** transfers effectively to standard CIR settings. To this end, we train a Standard CIR Baseline exclusively on the **OACIRR** training set and directly evaluate the resulting model in a zero-shot manner on three established CIR benchmarks: FashionIQ [18], CIRR [14], and CIRCO [4]. We compare its performance with representative CIR models trained on large-scale or synthetic triplet datasets, including CASE [9], CoVR-BLIP [16], CompoDiff [6], and CoAlign [10].

As shown in Table 5, the model pretrained on **OACIRR** achieves strong zero-shot transfer performance across all three benchmarks, consistently outperforming methods trained on substantially larger datasets. These findings support two key conclusions: **(1)** ***Importance of Instance-Consistent Supervision***: Enforcing precise instance-level alignment provides a more reliable training signal than synthetic or loosely paired semantic triplets, fostering robust compositional reasoning. **(2)** ***Data Efficiency through High Quality***: The real-world fidelity and careful curation of **OACIRR** lead to highly competitive transfer performance while requiring substantially fewer training samples than existing large-scale datasets. Overall, these cross-task results demonstrate that **OACIRR** serves not only as a rigorous benchmark for instance-aware retrieval, but also as an effective pretraining resource for the standard CIR task.

3.3. Cross-Domain Generalization on OACIRR

To evaluate whether models trained on **OACIRR** can generalize beyond domain-specific semantics, we conduct a leave-one-domain-out evaluation across the four subsets. For each target subset, the model is trained on the remaining three subsets and tested on the held-out one. We compare this **Cross-Domain** setting with the standard **Full Finetuning** setting, where all four subsets are used for training.

As shown in Table 6, ***AdaFocal*** consistently outperforms SPRC on all unseen domains under the Cross-Domain setting, demonstrating stronger instance-centric reasoning beyond domain-specific semantics. At the same time, the clear performance gap between Cross-Domain and Full Finetuning confirms that the four subsets are strongly complementary rather than redundant, highlighting both the diversity and the intrinsic challenge of the **OACIRR** benchmark.

3.4. Robustness to Bounding Box Quality

To evaluate the robustness of ***AdaFocal*** to imperfect user inputs, we simulate noisy bounding boxes through Scale and Shift perturbations. Specifically, *Scale* enlarges or shrinks the bounding box while preserving its center, and *Shift* additionally offsets the center to mimic localization errors.

As shown in Table 7, ***AdaFocal*** is robust to Scale pertur-

Bounding Box		Fashion			Car			Product			Landmark			Avg.
IoU	Perturbation	R_{ID}@1	R@1	R@5	R_{ID}@1	R@1	R@5	R_{ID}@1	R@1	R@5	R_{ID}@1	R@1	R@5	
1.00	***Original***	**77.15**	**65.31**	**86.88**	**78.42**	**53.63**	**92.22**	**91.86**	**74.11**	**95.39**	**82.92**	**58.47**	**91.63**	**79.00**
0.80	*Scale*	77.05	65.24	86.82	78.26	53.54	92.16	91.86	74.11	95.35	82.83	58.41	91.60	78.93
0.50	*Scale* + *Shift*	75.16	63.24	85.55	77.07	52.61	91.54	91.20	73.44	94.83	81.66	57.66	90.96	77.91
NaN	***w/o Bounding Box***	69.07	58.76	81.44	74.59	49.78	89.46	87.48	69.53	93.66	79.80	55.49	89.87	74.91

Table 7. Robustness of ***AdaFocal*** to Scale and Shift perturbations of bounding boxes on the **OACIRR** benchmark.

Modulation Output	Fashion			Car			Product			Landmark			Avg.
	R_{ID}@1	R@1	R@5	R_{ID}@1	R@1	R@5	R_{ID}@1	R@1	R@5	R_{ID}@1	R@1	R@5	
Scalar (β)	**77.15**	**65.31**	**86.88**	**78.42**	**53.63**	**92.22**	**91.86**	**74.11**	**95.39**	**82.92**	58.47	91.63	**79.00**
Vector ($\vec{\beta}$)	74.60	65.25	85.94	77.32	53.33	92.19	91.56	73.13	94.92	82.80	**58.96**	**91.77**	78.48

Table 8. Ablation study on the modulation output design of the *CAAM*.

CAAM	OACIRR Benchmark			
# Self-Attention Layers	R_{ID}@1	R@1	R@5	**Avg.**
1	81.38	62.39	90.54	78.10
2	**82.59**	**62.88**	**91.53**	**79.00**
3	82.31	62.75	91.42	78.83
4	82.02	62.51	91.24	78.59

Table 9. Ablation study on the number of *self-attention layers*.

CAAM	OACIRR Benchmark			
# Probe Tokens	R_{ID}@1	R@1	R@5	**Avg.**
2	81.92	**63.15**	91.49	78.85
4	82.38	62.41	91.15	78.65
8	**82.59**	62.88	**91.53**	**79.00**
16	82.46	62.94	91.45	78.95
32	82.21	62.90	91.37	78.83

Table 10. Ablation study on the number of *probe tokens*.

bation, with only negligible performance drops across all subsets. In contrast, the combined perturbation of *Scale* + *Shift* causes a clearer degradation, and removing the bounding box leads to the largest drop. These results indicate that ***AdaFocal*** tolerates moderate input noise while still relying on visual anchors for reliable instance-aware retrieval.

3.5. CAAM Design Analysis

We further analyze three key design choices of the *Context-Aware Attention Modulator (CAAM)*, including the modulation output form, the depth of the *Contextual Reasoning Module (CRM)*, and the number of learnable *Contextual Probe Tokens*. We prioritize configurations that achieve strong performance with minimal complexity.

- **Scalar vs. Vector Modulation.** As shown in Table 8, replacing the default scalar modulation with a query-wise vector output ($\vec{\beta} \in \mathbb{R}^M$) offers no additional gain. This suggests that a single scalar is sufficient to control attention intensity while better preserving the relative semantic coherence among pre-trained fusion queries. Therefore, we adopt the **scalar** design as the default output form.
- **Depth of the Contextual Reasoning Module.** As shown in Table 9, increasing the *CRM* depth from 1 to 2 layers leads to clear improvements in both instance-level fidelity and overall recall, indicating that a single layer lacks sufficient cross-modal reasoning capacity. Scaling beyond 2 layers offers no significant gains and may add unnecessary complexity to the compact design of the module. Based on these observations, we employ a **2-layer CRM**.
- **Number of the Contextual Probe Tokens.** As shown in Table 10, using too few probe tokens limits the module's capacity to capture diverse contextual cues, while increasing the token count further yields only marginal benefit. Since the performance saturates at **8 probe tokens**, we adopt this configuration as the default setting.

4. Additional Qualitative Analysis

Figure 2 presents qualitative comparisons across diverse retrieval scenarios, revealing two failure modes of baseline CIR models and showing how ***AdaFocal*** addresses them.

Semantic Drift. Baseline models tend to conflate strong textual modifications with intrinsic object attributes, yielding retrievals that follow text-implied properties rather than preserving the visual anchor. ***AdaFocal*** maintains instance identity while faithfully reflecting contextual changes.

Fine-grained Confusion. Baseline models often return semantically similar yet instance-incorrect distractors, reflecting a reliance on global semantics over instance-specific cues. ***AdaFocal*** retrieves the correct instance more reliably under high visual similarity, offering clear gains in challenging cases by emphasizing distinctive local cues.

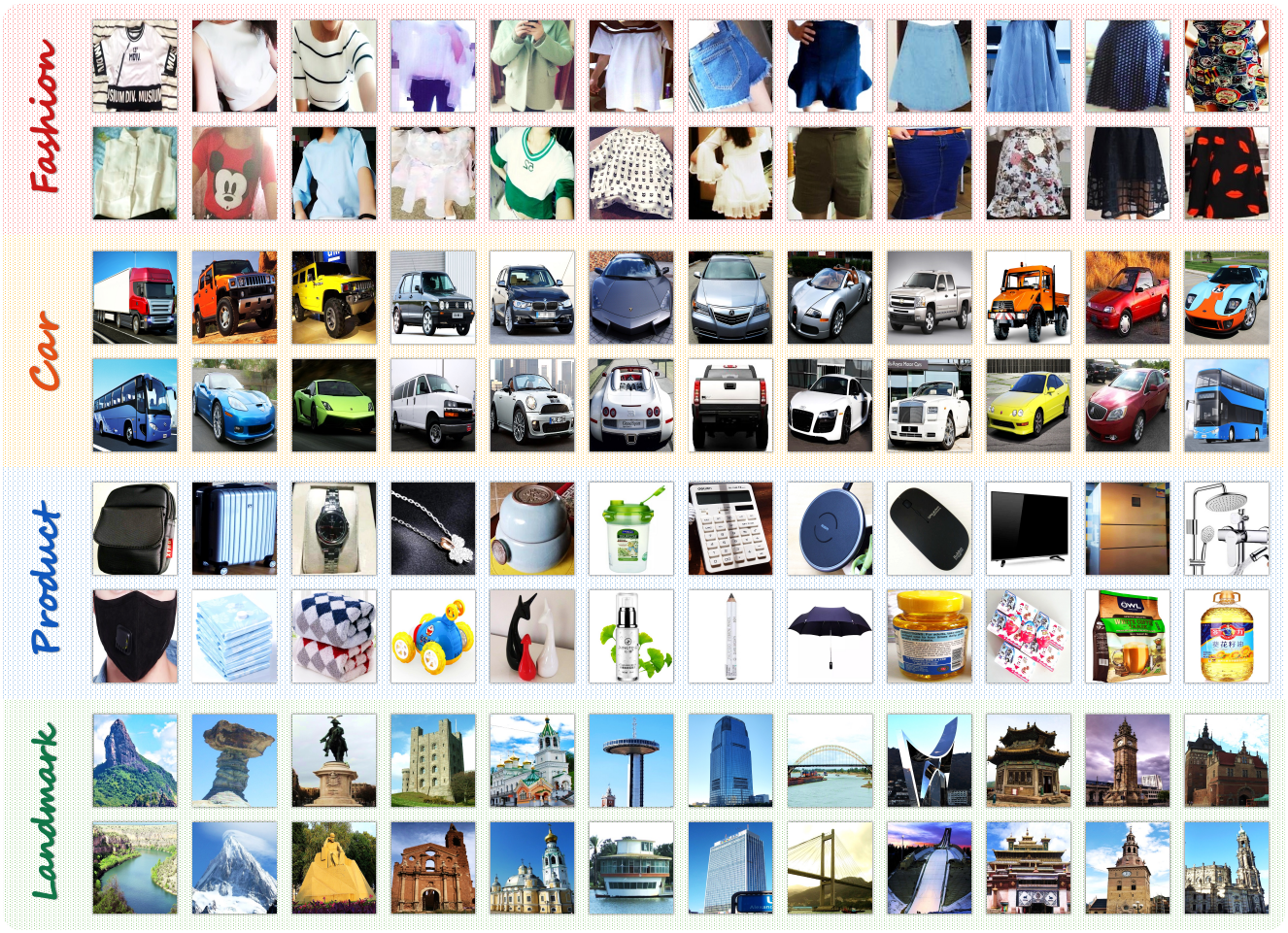

Figure 1. A curated collage of representative instances from the **OACIRR** benchmark.

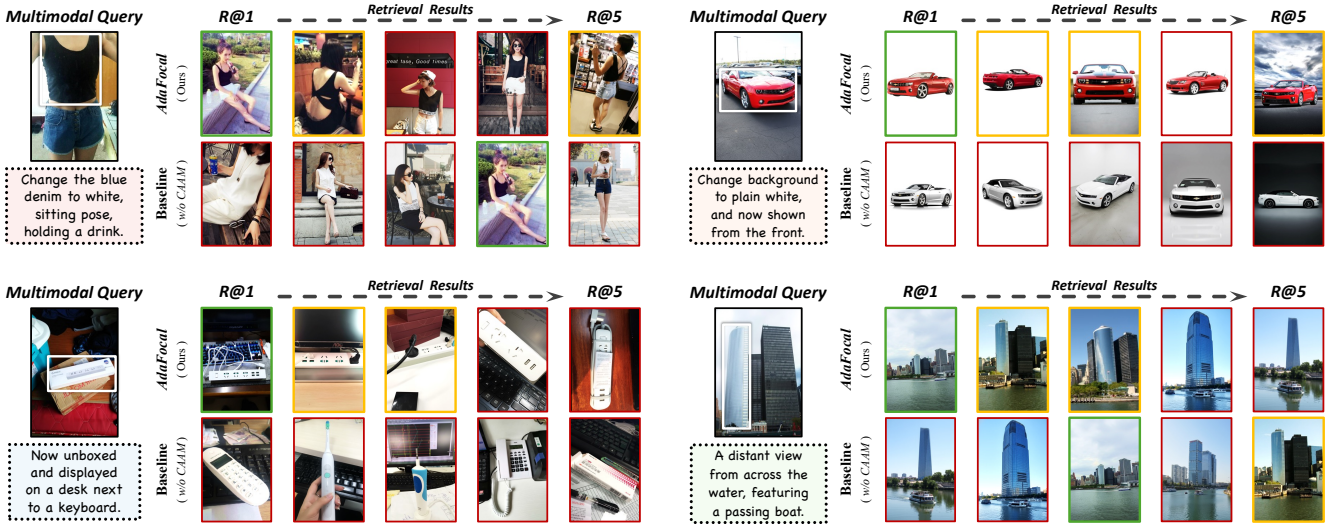

Figure 2. Qualitative comparison of our ***AdaFocal*** and the Baseline on the **OACIRR** benchmark. **Green boxes** indicate the ground-truth target, **yellow boxes** indicate instance-correct but semantically incorrect results, and all other retrieved images are marked with **red boxes**.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[2] Yalong Bai, Yuxiang Chen, Wei Yu, Linfang Wang, and Wei Zhang. Products-10k: A large-scale product recognition dataset. *arXiv preprint arXiv:2008.10545*, 2020. 1

[3] 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 *The Twelfth International Conference on Learning Representations*, 2024. 6, 7

[4] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15338–15347, 2023. 7

[5] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5337–5345, 2019. 1

[6] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *Transactions on Machine Learning Research*, 2024. Expert Certification. 7

[7] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13225–13234, 2024. 6

[8] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 1

[9] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2991–2999, 2024. 7

[10] Haiwen Li, Delong Liu, Zhaohui Hou, Zhicheng Zhao, and Fei Su. Automatic synthesis of high-quality triplet data for composed image retrieval. *arXiv preprint arXiv:2507.05970*, 2025. 7

[11] Xiaojie Li, Chu Li, Shi-Zhe Chen, and Xi Chen. U-marvel: Unveiling key factors for universal multimodal retrieval via embedding learning with mllms. *arXiv preprint arXiv:2507.14902*, 2025. 6

[12] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. In *The Thirteenth International Conference on Learning Representations*, 2025. 6

[13] Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4015–4025, 2025. 6

[14] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 7

[15] 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*, pages 19305–19314, 2023. 6

[16] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5270–5279, 2024. 7

[17] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2: A large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2575–2584, 2020. 1

[18] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11307–11317, 2021. 7

[19] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024. 6

[20] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024. 1