

---

# Instance-Level Composed Image Retrieval

## Supplementary Material

---

Bill Psomas<sup>1\*</sup> George Retsinas<sup>2\*</sup> Nikos Efthymiadis<sup>1</sup> Panagiotis Filntisis<sup>2,4</sup>  
Yannis Avrithis<sup>5</sup> Petros Maragos<sup>2,3,4</sup> Ondrej Chum<sup>1</sup> Giorgos Toliass<sup>1</sup>

<sup>1</sup>VRG, FEE, Czech Technical University in Prague    <sup>2</sup>Robotics Institute, Athena Research Center

<sup>3</sup>National Technical University of Athens    <sup>4</sup>Hellenic Robotics Center of Excellence    <sup>5</sup>IARAI

The supplementary material includes the following information and results:

- An ethics statement is presented in [section S1](#).
- The *i*-CIR structure, statistics and a sample of image and text queries are presented in [subsection S2.1](#).
- The shortcomings of some existing datasets (FashionIQ, CIRR, CIRCO) are discussed and supported by examples in [subsection S2.2](#).
- Additional technical and implementation details of BASIC are discussed in [section S3](#).
- Additional results are presented in [section S4](#). In particular:
  - The impact of hyper-parameter values in [subsection S4.1](#).
  - In [subsection S4.2](#) we show that the proposed components are effective beyond the scope of composed image retrieval, *i.e.*, for image-to-image and text-to-image retrieval too.
  - Detailed performance analysis per domain and category is presented in [subsection S4.3](#).
  - The performance comparison on instruction-like datasets is presented in [subsection S4.4](#).
  - The impact of corpora, and how they control the semantic projection, is further explored in [subsection S4.5](#).
  - Time comparison is presented in [subsection S4.6](#).
  - In [subsection S4.7](#) we present an experiment showing that *i*-CIR, despite being small, is as challenging as including more than 20M distractor images.
  - In [subsection S4.8](#) we present an experiment to demonstrate the *i*-CIR is not biased towards the model used in the data collection process.
  - In [subsection S4.9](#) we assess robustness to text-query wording by rewriting text queries, showing how phrasing alone can markedly shift retrieval performance.
  - In [subsection S4.10](#) we present retrieval results in *i*-CIR, visually comparing different components of BASIC.

## S1 Ethics Statement

### S1.1 Broader-Impact & Dual-Use Discussion

Instance-level composed image retrieval offers clear societal benefits: it can power fine-grained search across museum and GLAM archives, helping curators and the general public surface rare artifacts and provenance links that would be impractical to discover manually; it can also bolster assistive-vision tools that describe surroundings to blind or low-vision users by retrieving the precise

---

\*Equal contribution

objects referenced in natural-language queries. At the same time, the very same instance-level matching capabilities entail dual-use risks: in the wrong hands they could enable large-scale surveillance, doxxing, or targeted advertising by tracing a specific building, vehicle, or logo across web-scale image corpora.

Precisely because of these concerns, our dataset (*i*-CIR) was intentionally designed to minimize such potential misuse:

- *Privacy-first curation.* Annotators were trained to preferentially exclude images containing faces, license plates, private premises, or other personally identifiable information (PII) whenever doing so did not harm the task—*e.g.*, for landmark queries where abundant alternatives without PII existed. For categories where people are intrinsic to the depiction (*e.g.*, apparel modeled by humans), we retained the images but subsequently applied automatic, exhaustive anonymization: visible faces were pixelated across all retained images to protect identity while preserving the visual evidence needed for instance-level retrieval. Examples of all image queries and unique text queries are presented in [Figure S2](#) and [Figure S9](#) respectively.
- *Evaluation-only release.* *i*-CIR is published solely as an evaluation benchmark, not a training corpus. A model evaluated on *i*-CIR, containing no human-centric identifiers, cannot perform surveillance tasks without additional fine-tuning on sensitive data. While such re-purposing is possible in principle, it requires access to an external, privacy-violating dataset; our license explicitly forbids this.
- *Domain specificity.* Instance-level retrieval models are highly specialized to the visual domain on which they are trained. The gap between *i*-CIR (landmarks, products, fiction, fashion, tech gadgets) and human-centric surveillance domains further reduces direct transferability. Moreover, the categories of instances represented in *i*-CIR —landmarks, fictional characters, consumer products, fashion items, and technology gadgets (see [Figure S1](#))—are already common in benchmarks such as Google Landmarks (GLD) [16], INSTRE [15], Stanford Online Products [13], and In-Shop [9]; therefore, *i*-CIR does not introduce novel categories of high-risk.

## S1.2 Residual Dual-Use Vectors & Mitigations

- *Fine-tuning pathway – Risk:* Techniques (*e.g.*, architectures, loss functions) validated on *i*-CIR could later be fine-tuned on a face-centric or plate-centric corpus to build a surveillance model. *Mitigation:* Our CC-BY-NC-SA license forbids biometric or privacy-invasive applications, and downstream works must inherit these restrictions. We additionally provide a misuse policy and reserve the right to revoke access for violators. We recognize, however, that a determined adversary could ignore license terms. Our goal is to follow community best practice and exert every reasonable control that dataset creators can apply, while keeping the research benefits intact.
- *Dataset-construction pathway – Risk:* Our LAION-based search-and-retrieve scripts, if misused, could harvest a new dataset rich in faces or license plates. *Mitigation:* We release the scripts under the same restrictive license and with hard-coded filters that block sensitive keywords. The accompanying documentation explicitly instructs researchers not to re-purpose the pipeline for sensitive-content collection and encourages institutional review board (IRB) review for any modifications. Importantly, similar LAION-based search-and-retrieve scripts have previously been employed [2]; our code does not introduce a novel capability but re-implements such workflow with stricter safety defaults.
- *Object-of-interest tracking — Risk:* Even without identifying people directly, one could try to track an individual via their belongings (*e.g.*, a rare handbag or customized laptop sticker). *Mitigation:* We explicitly *prohibit* using *i*-CIR to identify, profile, or infer the movements or associations of any person—directly or indirectly via personal effects. Retrieval results from *i*-CIR must not be construed as evidence of co-location or identity. We will conduct periodic red-team evaluations focused on object-level re-identification risks and update the release if failure modes are found. The project page will prominently include a “Report misuse of the dataset” channel (web form/email) so the community can flag suspected abuse or PII leakage; reports will be reviewed promptly and may result in content takedown or revocation of access, consistent with our license.

**Report misuse.** *i*-CIR must not be used to identify, profile, or infer movements/associations of any person. If you believe the dataset or code is being used in such a way or you discover content that reveals PII—please submit a report via the “Report misuse of the dataset” form or email us. We acknowledge reports promptly and may take actions including content removal or access revocation under the license.

### S1.3 Transparency & Governance Commitments

- *License & API safeguards.* CC-BY-NC-SA with explicit surveillance prohibition; research-only API gated behind user agreements; automated checks to block bulk reverse-image searches of sensitive facilities.
- *Community reporting.* We provide clear channels for reporting misuse or ethical concerns (see “Report misuse” above). Substantiated reports will receive prompt acknowledgement and a public response timeline, and may result in content removal or access revocation consistent with the license.

### S1.4 Human-Annotation Protocol (Labor Conditions & Oversight)

- *Employment & Compensation.* All annotators are salaried employees of our institution (not crowd-workers). Their contracts include full social-security coverage and wages that exceed 80% of a first-year PhD stipend, comfortably above the legally mandated minimum in our region.
- *Training & Oversight.*
  - Domain training: A dedicated one-week training covered instance-level composed-image retrieval guidelines, annotation software, and quality-control protocols.
  - Ethics training: The same program included refresher modules on copyright compliance, privacy protection, and sensitive-content redaction.
  - Ongoing monitoring: Senior staff (including authors) conducted random spot-checks and weekly inter-annotator-agreement audits; annotators receive feedback whenever discrepancies arise, and, if needed, re-labeling was performed.

## S2 More on *i*-CIR and existing datasets

### S2.1 Structure, statistics, visualisations, and limitations of *i*-CIR

**Taxonomies.** To better understand the diversity and structure of *i*-CIR, we organize all object instances into a 3-level hierarchy of visual categories and all textual modifications into a 1-level taxonomy of textual categories.

The visual taxonomy captures the type and nature of the object instance depicted in the image query. It begins with broad classes such as landmark, fictional, product, tech, and mobility, which are further refined into subcategories and specific object types. For example, “Asterix” is categorized as fictional → character → comic, while “Temple of Poseidon” falls under landmark → architecture → temple.

In parallel, each text query is annotated with one of seven high-level textual categories based on the nature of the transformation it describes:

- *Addition:* One or more external elements are introduced into the scene alongside the instance — such as people, objects, or animals. Examples: “with a man proposing”, “next to coffee beans”.
- *Appearance:* A full transformation of the instance’s physical form or structure. The object is still the main focus but appears in a completely different embodiment. Examples: “as a figurine”, “as a sculpture”, “as a scale model”.
- *Attribute:* A partial modification of the instance itself — such as color, minor structural differences, or subtle material changes. The identity remains unchanged. Examples: “in purple color”, “with yellow shoelaces”.
- *Context:* A modification in the surrounding environment or scene, lighting conditions, or time of day. These affect the setting without altering the object instance itself. Examples: at night, during sunset, outdoors on grass.

- *Domain*: The instance is rendered in a different representational domain or medium, such as sketches, paintings, 3D renders, magazine ads, or comics. Examples: “as a painting”, “in a manga panel”.
- *Projection*: The object instance is placed onto another object or surface, often as decoration or branding — such as clothing, packaging, or household items. Examples: “on a t-shirt”, “printed on a pillow”.
- *Viewpoint*: A change in camera/viewer perspective, such as aerial shots or top-down views. The instance and environment remain the same. Examples: “from an aerial viewpoint”, “from a top-down viewpoint”.

These taxonomies, illustrated in [Figure S1](#), provides insight into the types of visual and textual variation covered in *i*-CIR and supports performance breakdowns by category.

**The 202 instances.** [Figure S2](#) provides a visual overview of the 202 object instances included in *i*-CIR. We intentionally set the total to 202—200 general instances plus 2 pet/animal instances (dogs)—included as a symbolic nod to the pets of our team while keeping the benchmark broad. Each image corresponds to a distinct instance, ranging from iconic landmarks and branded products to fictional characters and vehicles. Roughly half of the instances are *nameable* (with a canonical, widely recognized name; *e.g.*, “Eiffel Tower”), while the remaining half are everyday objects that are best described compositionally (*e.g.*, a “white-and-pink dolphin plushie”). All main-paper results are reported on the full *i*-CIR; in ablations below we also report on two subsets: *i*-CIR<sub>named</sub> (only nameable instances) and *i*-CIR<sub>descriptive</sub> (only compositionally described instances).

**Unique text queries.** [Figure S9](#) presents the complete list of unique textual queries used in *i*-CIR. Each query describes a specific modification applied to an object instance, ranging from appearance and material to setting, domain, or interaction. These compositional cues form the basis of our retrieval queries and illustrate the rich semantic diversity captured in the benchmark.

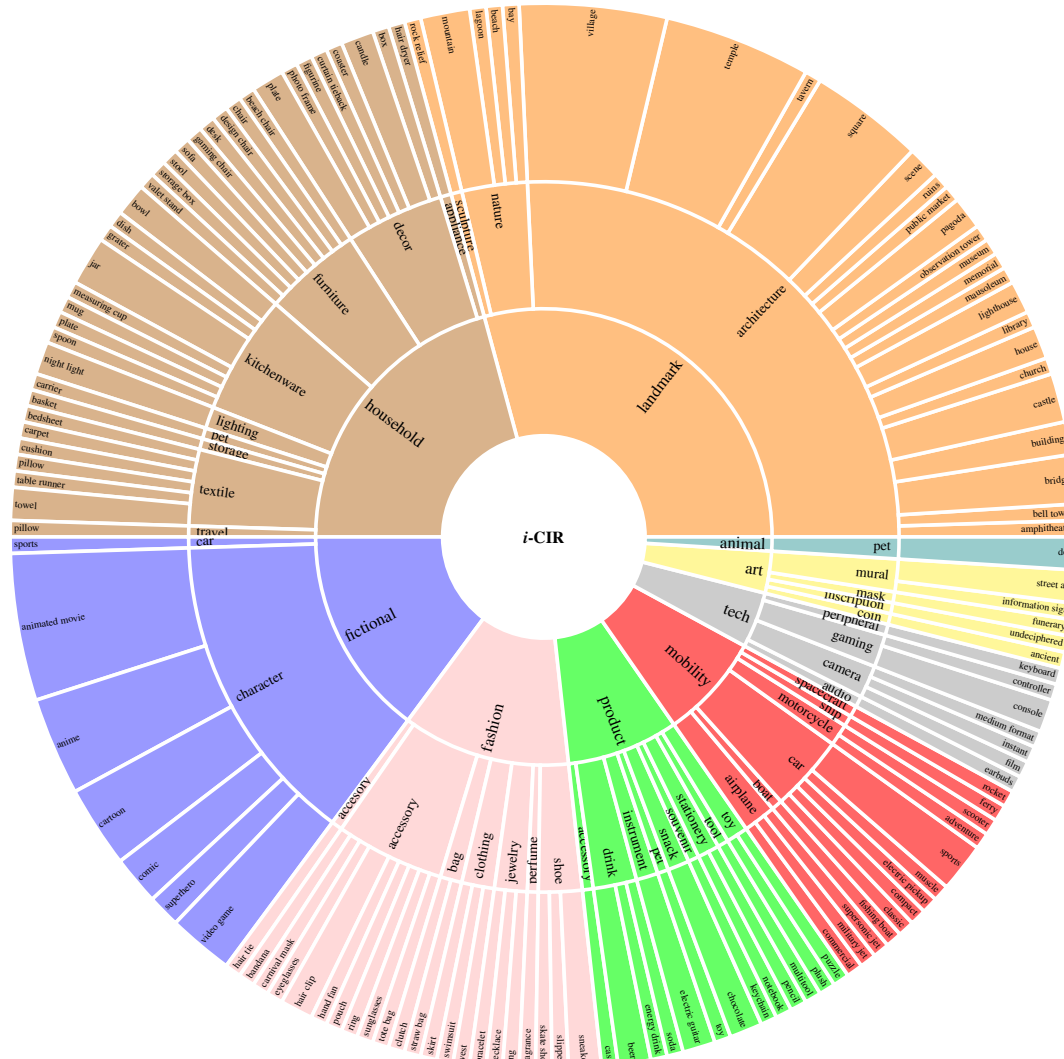
**Personally identifiable information redaction (face pixelation).** To protect privacy while preserving task utility, we automatically *pixelated* visible human faces across *i*-CIR. We used off-the-shelf detectors to localize faces and required overlap with a person detector before redaction, which reduces false positives from posters, figurines, or mannequins; detected boxes were conservatively expanded before applying a mosaic filter. For categories where people are incidental (*e.g.*, landmarks), such images were preferentially filtered by annotators during curation; where people are intrinsic to the content (*e.g.*, apparel), we retained the images but anonymized all faces exhaustively. Thresholds were set for high recall and audited via stratified spot checks; license plates and other PII were similarly removed or obfuscated when encountered.

**Limitations of *i*-CIR** Our semi-automatic pipeline trades some speed for fidelity: every composed query is manually vetted (*e.g.*, for hard negatives, PII issues, etc.), so building *i*-CIR is slower and costlier than fully automatic datasets, but yields an ambiguity-free benchmark the field currently lacks. Moreover, the design mostly assumes a single salient object per image query (see [Figure S2](#)) and uses English-only phrasing with public-domain imagery. Despite filtering and automated redaction, some residual PII may persist; based on stratified manual audits of random samples across categories, we estimate the prevalence of unredacted faces or other PII to be < 2%. Our project page will include a “Report misuse / PII” channel, and we commit to prompt review and content updates or takedown upon verified reports.

## S2.2 Shortcomings of existing benchmarks

**CIRR.** To substantiate our claims regarding the limitations of existing CIR benchmarks, we present qualitative retrieval results from the CIRR dataset using Text  $\times$  Image in [Figure S10](#). In “fewer paper towels per pack”, a semantically relevant match appears as the third result, yet it is incorrectly annotated as a negative, showcasing a clear false negative. In “the target is a Pepsi bottle”, the query text alone suffices to retrieve relevant images containing Pepsi bottles; the image query adds minimal value beyond indicating the Pepsi brand. Notably, the fifth and eighth retrieved images include visible Pepsi bottles, but are labeled as negatives—likely due to the co-occurrence of other non-Pepsi items, further exposing annotation inconsistencies. In “two





(a) *Visual categories* organized in a 3-level hierarchy.



(b) *Textual modification categories* labeled at a single level.

Figure S1: *The i-CIR taxonomies* showing (a) the 3-level hierarchy of visual categories and (b) the single-level taxonomy of textual modification categories. This taxonomy captures the diversity and distribution of object instances and textual queries, and enables performance reporting per category.

antelopes standing one in front of the other”, again, the textual query alone retrieves the intended concept, rendering the image query unnecessary for composition. Yet, the seventh result clearly satisfies the query but is misclassified as a negative. Lastly, in “widen the rows of TVs”, the supposed ground-truth positive contains four rows of TVs—just as in the image query—while the sixth retrieved image includes five rows, arguably an even better match. Nevertheless, it is annotated as a negative, underlining the prevalence of false negatives and the lack of fine-grained visual reasoning in the benchmark.

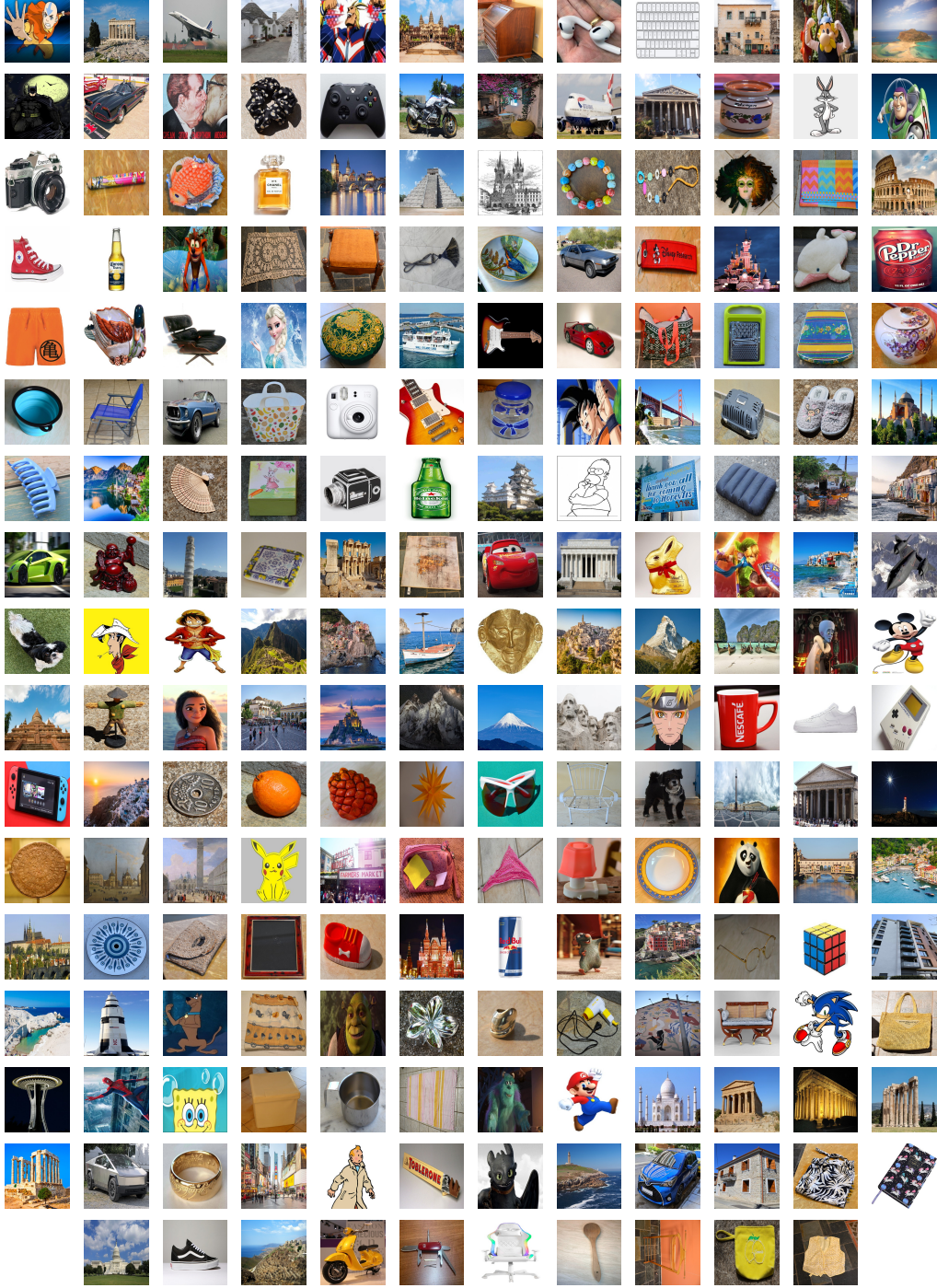


Figure S2: The 202 instances of *i*-CIR dataset. We randomly sample a query image per object instance.

**FashionIQ.** Figure S11 visualizes retrieval results on FashionIQ using Text  $\times$  Image, revealing several inconsistencies. In many cases, such as “is solid black, is long” or “is green with a four leaf clover, has no text”, correct results appear within the top-10 but are annotated as negatives, showing false negative noise. In other cases, the query image plays a minimal or redundant role; for instance, “is the same” easily retrieves the target based on visual matching alone, reducing the task to a simpler image-only similarity search. These observations support our

claim that FashionIQ, like other CIR benchmarks, suffers from supervision that stems from retrofitting relations between pre-existing images. As a result, the composed queries do not always require both modalities to be resolved.

**CIRCO.** Figure S12 presents retrieval examples from CIRCO using Text  $\times$  Image, which appears to have less flaws than CIR and FashionIQ. While CIRCO exhibits fewer problematic cases, it is not entirely free of issues. Due to the class-level nature of the dataset, the distinction between positive and negative samples can sometimes be ambiguous—even for human annotators. For instance, in the case of the query “shows only one person”, visually plausible matches like the third retrieved image clearly align with the query yet are labeled as negatives. Similarly, for the query “has more of them and is shot in greyscale”, multiple retrieved images fulfill both criteria but remain unannotated as positives. These examples highlight the challenge of assigning unambiguous relevance labels at the class-level, even in well-curated datasets.

### S3 More on BASIC

**On projection matrix  $\mathbf{P}$ .** As described in the main manuscript (Sec. 4), we form the projection matrix  $\mathbf{P} \in \mathbb{R}^{d \times k}$  by using the top- $k$  eigenvectors of  $\mathbf{C}$ :

$$\mathbf{C} = (1 - \alpha) \mathbf{C}_+ - \alpha \mathbf{C}_-, \quad \text{where} \quad \mathbf{C}_\pm = \frac{1}{|C_\pm|} \sum_{x \in C_\pm} (\phi^t(x) - \boldsymbol{\mu}^t)(\phi^t(x) - \boldsymbol{\mu}^t)^\top.$$

These eigenvectors capture directions of high variance in  $\mathbf{C}_+$  and low variance in  $\mathbf{C}_-$  when their corresponding eigenvalues are positive. Conversely, eigenvectors associated with negative eigenvalues indicate directions of high variance in  $\mathbf{C}_-$  and low variance in  $\mathbf{C}_+$ , which we do not wish to retain.

To address this, we first count the number of “useful” components,

$$k_+ = \#\{\lambda_i > 0\},$$

*i.e.*, the number of eigenvalues of  $\mathbf{C}$  that are positive. We then set the final number of components as

$$k \leftarrow \min(k, k_+).$$

In practice, for moderate values of the mixing parameter  $\alpha$ , the initial choice of  $k$  (*i.e.*,  $k = 250$ ) typically already excludes negative eigenvalues. However, for large  $\alpha$  (*e.g.*,  $\alpha = 0.8$ ) negative eigenvalues can appear, necessitating this refinement.

**Query expansion details.** Given the initial projected query image embedding  $\mathbf{P}^\top \bar{\mathbf{q}}^v \in \mathbb{R}^d$ , we compute its top- $k$  nearest neighbors  $\{z_1^v, \dots, z_k^v\} \subset X$  in the database  $X$ . This set is augmented with the query image itself, to avoid deviating from the original query. The centered features  $\bar{\mathbf{z}}_i^v = \phi^v(z_i^v) - \boldsymbol{\mu}^v \in \mathbb{R}^d$  are then aggregated into an expanded query  $\tilde{\mathbf{q}}^v \in \mathbb{R}^d$  via a soft attention-weighted mean:

$$\tilde{\mathbf{q}}^v = \sum_{i=1}^{k+1} w_i \bar{\mathbf{z}}_i^v, \quad \text{where} \quad w_i = \frac{\exp(\beta \langle \mathbf{P}^\top \bar{\mathbf{z}}_i^v, \mathbf{P}^\top \bar{\mathbf{q}}^v \rangle)}{\sum_{j=1}^{k+1} \exp(\beta \langle \mathbf{P}^\top \bar{\mathbf{z}}_j^v, \mathbf{P}^\top \bar{\mathbf{q}}^v \rangle)}. \quad (\text{S1})$$

The “temperature” hyper-parameter  $\beta$  is fixed to 0.1. Again, following the computational complexity analysis of Sec. 4, we can avoid operating directly on the dataset features. In particular, the top- $k$  retrieval can be accelerated (*e.g.*, using FAISS [5]) by the formulation in the main manuscript, where sorting can be done using only the term  $\langle \mathbf{x}^v, \mathbf{P} \mathbf{P}^\top (\mathbf{q}^v - \boldsymbol{\mu}^v) \rangle$ .

**Contextualization insights.** Contextualization provides a notable boost across all datasets (Table 2, main manuscript). However, the extent of this improvement varied by dataset. For example, contextualization does not produce the same gains for NICO++ as does for others. This dataset includes domains like “ock”, “water”, and “outdoors”, where combining them with objects (*e.g.*, “rock dog” or “dog rock”) does not result in a particularly meaningful representation. In contrast, datasets like *i*-CIR benefit more clearly from context, *e.g.*, “dog during sunset”.

In some cases, adding the right level of textual detail, such as “a dog in a rocky environment”, is important for forming a well-defined text query. This can be crucial for further unlocking the

potential of BASIC. However, generating such enriched text requires the use of LLMs, which can introduce prohibitive overhead when applied on-the-fly for each query. Alternatively, a lightweight (shallow) network trained specifically to contextualize abstract queries could be used directly on their embedding space. While this is a promising direction, it is not explored in this work.

**Corpora creation.** To create the various corpora used in this work, we used ChatGPT. We provided a carefully designed prompt with few examples and generated batches of entries (e.g. 100 corpus entries per prompt). Typically, the created corpus needs a post-processing of removing duplicates and erroneous outputs. For example, there were cases that after a number of entries, ChatGPT returned noisy data, such as Description#87, Description#88 etc.

In particular, the prompt used to create the generic object corpus was:

**Generic object corpus prompt**

I would like you to generate a vocabulary of roughly 2,000 visual classes that are similar in style and granularity to the categories found in ImageNet-1K. These should include:

- Objects (e.g., “espresso maker”, “screwdriver”)
- Animals (e.g., “Siberian husky”, “hummingbird”)
- Scenes or natural elements (e.g., “iceberg”, “volcano”)
- Everyday items (e.g., “shopping cart”, “paper clip”)
- Tools, instruments, clothing, vehicles, and so on.

The classes should cover a broad range of everyday and recognizable visual categories, suitable for training or evaluating vision-language or classification models. Some class names may be single-word (e.g., “zebra”) while others may be multi-word expressions (e.g., “motor scooter”, “garden hose”).

Please do not include brand names or highly specialized terms. Favor categories that have visually distinguishable appearances.

Output the vocabulary in chunks of 100 class names at a time, formatted as a plain list (one per line). Wait for me to say “next” before giving the next chunk.

In total, the generic corpora used consist of 1,800 entries in the object corpus and 1,000 stylistic elements in the negative corpus. To create the negative word corpus, which captures stylistic and contextual variations, we used the following prompt:

**Negative word corpus prompt**

We are working on a composed image retrieval task, where the input consists of an image and a text. The goal is to retrieve the most relevant matching image from a database. Typically, the image contains a main object in a particular “state”. This state can refer to a style (e.g., retro, handwritten, futuristic) or a contextual setting (e.g., next to the sea, on the beach, beside a table). For example: a retro mug, a cat next to a table, a handwritten notebook. We would like you to generate a .txt file containing 100 possible states in which an object can be found. Each line in the file should contain a single state expressed in natural language. Wait for me to say “next” before giving the next batch of states.

The first outputs of the above prompt are presented in [Figure S3](#).

**Min normalization.** To normalize and balance the two query modalities, we proposed min-based normalization. To perform this step, the corresponding statistical values  $s_{\min}^v$  and  $s_{\min}^t$  must be



- on a wooden shelf
- painted with graffiti
- surrounded by candles
- floating in water
- next to a fireplace
- hanging from a tree
- wrapped in plastic
- sitting on grass
- under a spotlight

Figure S3: *Examples from the generic negative corpus.* Entries are natural-language style and contextual setting modifiers.

precomputed. In other words, we empirically estimate lower bounds on image-to-image and text-to-image similarities after the centralization step. These values can be extracted from a large existing dataset (e.g., LAION [12]). Instead, we opted to use a small synthetic dataset, generated via Stable Diffusion [11], where an image is created directly from its text description.

There are two main reasons for this choice: 1) we obtain an accurate correspondence between each text caption and its generated image, and 2) we compute all pairwise similarity scores (which scale quadratically with dataset size), allowing us to efficiently recompute statistics under different settings (e.g., with or without applying the projection step). Performing this over millions of embeddings would be prohibitive, whereas the required statistics are conceptually simple values.

Specifically, we first used ChatGPT to generate 100 simple and diverse textual prompts (e.g., ‘‘a dog’’, ‘‘a calm sea’’) and 100 more complex ones (e.g., ‘‘a hidden waterfall in a lush, green jungle’’). Each prompt was then used to generate 4 corresponding images using Stable Diffusion [11], encouraging intra-prompt diversity. All image-text and image-image pairs were then processed through our pipeline using the same CLIP backbone, and their similarity scores were computed. These scores were used to derive the normalization statistics  $s_{\min}^v$  and  $s_{\min}^t$ , which are calculated after centralization.

For normalization, the minimum similarity values should always be negative. This is expected, as the pipeline is built upon cosine similarity, and given a sufficiently diverse set, centralization alone cannot shift all similarities into the positive range. This behavior is empirically validated in our experiments. Indicative values for the CLIP model are:  $s_{\min}^v = -0.077$  and  $s_{\min}^t = -0.117$ .

**Limitations of BASIC.** BASIC assumes that both image and text components of a query provide semantically distinct and complementary, equally important, information. This allows us to interpret similarity multiplication as a logical conjunction operator. However, in tasks where one modality dominates (e.g., text-dominated edits as in FashionIQ [17] and CIRR [8]; examples shown in Figure S11 and Figure S10 respectively) or where the text query is highly entangled with image content (e.g., CIRCO [1] text queries like ‘has a dog of a different breed and shows a jolly roger’), this assumption may not hold. In such cases, the benefits of our projection and fusion mechanism may diminish, or even have a negative impact, compared to using only the textual modality.

## S4 More experiments

### S4.1 Hyperparameter search of BASIC

During the development of BASIC, we set aside a small portion of the *i*-CIR crawl as a development set; none of these images appears in the final test benchmark. *i*-CIR<sub>dev</sub> consists of 15 object instances, 92 composed queries, and 45K images in total. All BASIC hyperparameters: contrastive scaling  $\alpha$ , number of PCA components  $k$ , and Harris regularization  $\lambda$ , were tuned once on this split and then frozen. To test the robustness of these choices, Table S1 runs a *leave-one-dataset-out* check: we treat each dataset in turn as the development set (re-tuning  $\alpha$ ,  $k$ ,  $\lambda$  on that dataset only) and then evaluate the resulting configuration on all datasets. As expected, the best score per row typically occurs on the diagonal (in-domain), but the off-diagonal performance is very close, indicating that BASIC is not overly sensitive to where the hyperparameters are tuned. Notably, the configuration tuned on *i*-CIR<sub>dev</sub> performs on par with the cross-dataset alternatives; *i*-CIR<sub>dev</sub> is a small convenience split used during development and is not part of the public release.

To further assess the sensitivity of BASIC to its hyperparameters, Figure S4 presents one-factor-at-a-time sweeps over the contrastive scaling  $\alpha$ , the number of PCA components  $k$ , and the Harris regularization weight  $\lambda$ , holding the other two fixed. Across all datasets, the curves exhibit a broad

Table S1: *Cross-dataset hyperparameter stability for BASIC*. Each row lists the hyperparameters selected when tuning on the “Dev” dataset; each column reports the resulting score when evaluating that configuration on the “Test” dataset (higher is better). Bold indicates in-domain evaluation.

Dev → Test	<i>i</i> -CIR	ImageNet-R	NICO++	MiniDN	LTLL	Avg.
<i>i</i> -CIR ( $\alpha = 0.2, k = 300, \lambda = 0.05$ )	<b>32.95</b>	30.60	30.21	38.50	41.02	34.66
ImageNet-R ( $\alpha = 0.4, k = 200, \lambda = 0.1$ )	30.67	<b>32.63</b>	31.39	38.49	40.41	34.72
NICO++ ( $\alpha = 0.2, k = 200, \lambda = 0.1$ )	30.95	32.55	<b>31.85</b>	39.81	41.32	35.30
MiniDN ( $\alpha = 0.2, k = 200, \lambda = 0.1$ )	30.95	32.55	31.85	<b>39.81</b>	41.32	35.30
LTLL ( $\alpha = 0.0, k = 300, \lambda = 0.1$ )	32.42	28.96	30.16	37.65	<b>42.44</b>	34.33
<i>i</i> -CIR <sub>dev</sub> ( $\alpha = 0.2, k = 250, \lambda = 0.1$ )	31.64	32.13	31.65	39.58	41.38	35.28

plateau around the optimum: while the exact maximizer may shift slightly within the (coarse) grid, it remains in the same neighborhood. We observe no sharp instabilities—moderate deviations from the tuned values incur only gradual changes in mAP—and the qualitative trends are consistent across datasets. Taken together, these results indicate that BASIC is robust to hyperparameter selection and that a single configuration transfers well across benchmarks.

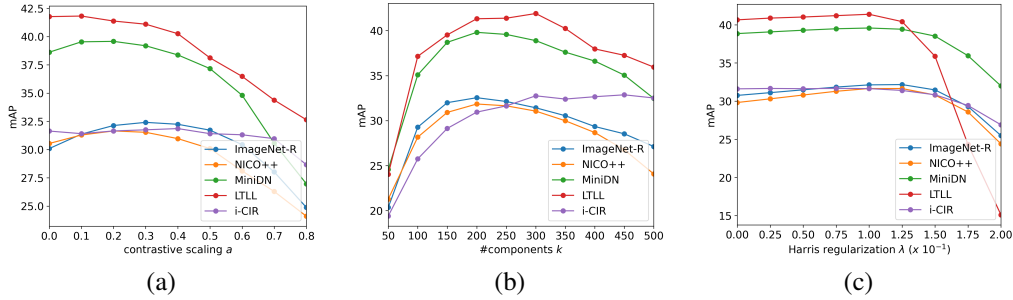


Figure S4: *Sensitivity of BASIC to hyperparameters*. mAP (%) on each dataset when varying one hyperparameter and holding the other two fixed: (a) contrastive scaling  $\alpha$ , (b) number of PCA components  $k$ , and (c) Harris regularization weight  $\lambda$ . Lines correspond to datasets in the legend (higher is better).

## S4.2 Unimodal evaluation of BASIC components

To better understand the contribution of each component in BASIC, we conduct a unimodal evaluation, by analyzing the system’s performance when querying with only one modality at a time. As shown in Table 2 of the main manuscript, the components of BASIC work in unison to produce the reported overall improvement. However, isolating the effects of each component in a unimodal setting can offer valuable insights into their individual contributions.

In this experiment, we consider two separate retrieval tasks: retrieving the most relevant images given a *text* query, and retrieving the most relevant images given an *image* query. We choose ImageNet-R for this analysis, as it allows for a clean decomposition of the two modalities. Table S2 presents the results, including a per-domain breakdown.

The following *key observations* can be made, supporting the analysis of the main manuscript:

- *Centering* consistently improves performance for both text and image queries.
- *Contextualization* significantly boosts performance in the text modality, with an improvement of over 6% mAP, highlighting its importance in enriching sparse or abstract textual queries.
- The *semantic projection step* offers the most substantial gain, increasing mAP by over 17%, underscoring the importance of content-only preservation in the image modality.
- *Query expansion* further enhances the effect of projection, adding an additional 3% improvement in mAP, acting as a complementary refinement mechanism. Notably, in the ORIGAMI case, this step reduces performance.

Table S2: Unimodal performance comparison in terms of mAP (%) using CLIP [10] for ImageNet-R. For each source domain (columns), we report the average mAP across all corresponding target domains. Each row progressively adds components in each modality, starting from the text-only or image-only baselines, respectively.

Modality	Method	CAR	ORI	PHO	SCU	TOY	AVG
textual	Text	67.02	67.00	74.94	71.77	77.31	71.61
	+ centr.	72.12	72.52	79.62	76.10	79.53	75.98
	+ context.	79.73	79.57	81.42	85.21	84.76	82.14
visual	Image	28.62	19.77	47.46	29.40	33.80	31.81
	+ centr.	35.04	21.64	50.87	32.72	38.24	35.70
	+ proj.	55.70	37.23	64.93	52.37	55.02	53.05
	+ q. exp.	59.17	36.43	67.89	56.73	59.90	56.02

### S4.3 Detailed comparisons across domains and categories

To further understand how BASIC performs across different domains in the evaluated datasets, as well as across the visual and textual categories of *i*-CIR, we provide detailed mAP results in Table S3 and Table S4, using two different vision-language models: CLIP [10] and SigLIP [18].

**CLIP.** Table S3 presents detailed results across domains and categories using CLIP [10] as backbone. Several key observations emerge:

- BASIC consistently outperforms all other approaches in the majority of domains and categories.
- FreeDom is the only method that achieves better performance than BASIC in any domain. Notably, this occurs only in the source domain PHOTO, and only within two datasets: ImageNet-R and MiniDN.
- In ImageNet-R, the source domain ORIGAMI is particularly challenging for all methods. For BASIC, ablating the negative-corpus term ( $\alpha=0$ ) drops mAP from 21.11% to 16.65%, whereas emphasizing it ( $\alpha=0.4$ ) raises mAP to 23.20%. This highlights how certain stylistic elements (e.g., “origami”) can become entangled with the underlying semantic content, and how the contrastive formulation of the corpora helps disentangle them.
- On domain–conversion datasets, the ranking is stable: BASIC is consistently first and FreeDom a reliable second. On *i*-CIR, the runner-up is category-dependent: MagicLens is most often second and, in the *household* and *fashion* visual categories, it even ranks first—likely reflecting its training distribution (consumer/e-commerce imagery). Despite these fluctuations, BASIC remains the top method overall and the best performer across the majority of visual and textual categories.

**SigLIP.** Table S4 presents detailed results using SigLIP [18] as backbone. Not all methods in our evaluation are compatible with the SigLIP backbone. Approaches such as MagicLens and Pic2Word rely on training lightweight models (e.g., MLPs or small Transformers) directly on top of CLIP features, making them inherently tied to CLIP’s representation space. Since SigLIP does not offer compatible pre-trained heads for these methods and retraining them on SigLIP features falls outside our scope, we restrict our comparison to FreeDom—a training-free method that directly operates on image and text features from the backbone. The following observations can be made:

- SigLIP improves overall retrieval performance across most datasets, with the only exception being NICO++, where a drop in performance was reported.
- BASIC significantly outperforms FreeDom on ImageNet-R, LTLL, and *i*-CIR. However, it reports lower mAP on both MiniDN and NICO++. Interestingly, in MiniDN, the two methods show complementary behavior—each outperforming the other on roughly half of the categories. This suggests that combining the strengths of both methods may further improve performance.
- In *i*-CIR, there are categories where the *Text*  $\times$  *Image* baseline is on par or surpasses FreeDom. This highlights FreeDom’s limitations in adapting to fine-grained, instance-level retrieval.



Table S3: Performance comparison in terms of mAP (%) using CLIP [10] across five datasets. (a-d): four domain conversion, where for each source domain (columns), we report the average mAP across all corresponding target domains; (e-f): *i*-CIR, where AP performance is averaged over queries grouped by their respective visual (e) or textual (f) category. AVG: average mAP over all source-target domain combinations. **Bold**: best; **purple**: second best.

Method	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.75	0.66	0.60	0.80	0.75	0.71
Image	3.89	3.53	1.02	6.03	5.82	4.06
Text + Image	5.89	5.00	2.66	9.23	9.18	6.39
Text $\times$ Image	7.22	5.72	7.49	8.65	9.19	7.66
Pic2Word	7.60	5.53	7.64	9.39	9.27	7.88
CompoDiff	13.71	10.61	8.76	15.17	16.17	12.88
WeiCom	10.07	7.61	10.06	11.26	13.38	10.47
SEARLE	18.11	9.02	9.94	17.26	15.83	14.04
MagicLens	7.79	6.33	11.02	9.94	10.57	9.13
FreeDom	<b>35.97</b>	<b>11.80</b>	<b>27.97</b>	<b>36.58</b>	<b>37.21</b>	<b>29.91</b>
BASIC	<b>36.49</b>	<b>21.11</b>	<b>26.07</b>	<b>39.30</b>	<b>37.68</b>	<b>32.13</b>

(a) ImageNet-R

Method	CLIP	PAINT	PHO	SKE	AVG
Text	0.63	0.49	0.62	0.50	0.56
Image	8.19	7.42	5.12	7.52	7.06
Text + Image	11.08	9.74	10.71	8.20	9.94
Text $\times$ Image	9.72	7.21	15.55	5.44	9.48
Pic2Word	13.39	8.63	17.96	8.03	12.00
CompoDiff	19.06	24.27	23.41	25.05	22.95
WeiCom	7.52	7.04	15.13	4.40	8.52
SEARLE	25.04	18.72	23.75	19.61	21.78
MagicLens	24.40	17.54	28.59	9.71	20.06
FreeDom	<b>41.96</b>	<b>31.65</b>	<b>41.12</b>	<b>34.36</b>	<b>37.27</b>
BASIC	<b>46.44</b>	<b>34.84</b>	<b>38.70</b>	<b>38.34</b>	<b>39.58</b>

(b) MiniDomainNet

Method	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Text	1.05	1.03	1.14	1.25	1.13	1.05	1.11
Image	6.82	5.00	6.02	8.02	8.06	5.91	6.64
Text + Image	8.99	6.89	9.66	12.34	11.91	8.61	9.74
Text $\times$ Image	7.99	6.32	11.40	11.64	10.09	8.11	9.26
Pic2Word	9.79	8.09	11.24	11.27	11.01	7.16	9.76
CompoDiff	10.07	7.83	10.53	11.41	11.93	10.15	10.32
WeiCom	8.58	7.39	13.04	13.17	11.32	9.73	10.54
SEARLE	13.49	13.73	17.91	17.99	15.79	11.84	15.13
MagicLens	18.76	15.17	22.14	23.61	21.99	16.30	19.66
FreeDom	<b>24.35</b>	<b>24.41</b>	<b>30.06</b>	<b>30.51</b>	<b>26.92</b>	<b>20.37</b>	<b>26.10</b>
BASIC	<b>31.06</b>	<b>31.07</b>	<b>34.62</b>	<b>36.69</b>	<b>32.18</b>	<b>24.24</b>	<b>31.65</b>

(c) NICO++

Method	TODAY	ARCHIVE	AVG
Text	5.09	5.30	5.20
Image	9.06	23.60	16.33
Text + Image	10.17	24.47	17.32
Text $\times$ Image	15.88	23.67	19.78
Pic2Word	17.86	24.67	21.27
CompoDiff	15.45	27.76	21.61
WeiCom	24.56	28.63	26.60
SEARLE	20.82	30.10	25.46
MagicLens	<b>33.77</b>	14.65	24.21
FreeDom	30.95	<b>35.52</b>	<b>33.24</b>
BASIC	<b>42.76</b>	<b>40.00</b>	<b>41.38</b>

(d) LTLL

Method	FICT	LAND	MOB	HOUSE	TECH	FASH	PROD	ART
Text	4.79	3.91	2.30	3.48	1.33	2.30	1.58	0.76
Image	1.05	2.43	1.78	2.64	2.52	2.18	1.82	1.77
Text + Image	2.56	12.85	8.18	3.96	9.92	3.48	4.18	4.86
Text $\times$ Image	21.05	27.88	25.50	11.25	22.97	11.87	18.10	26.81
Pic2Word	19.38	32.85	21.27	13.95	14.18	14.56	15.39	33.54
CompoDiff	10.51	15.06	19.27	3.56	6.18	7.41	4.54	13.04
WeiCom	18.15	30.21	20.27	8.82	23.30	12.22	18.67	34.27
SEARLE	<b>31.10</b>	31.05	18.52	11.69	16.97	14.39	17.23	22.75
CIReVL	28.66	29.64	23.39	14.06	<b>23.75</b>	11.74	24.89	19.04
MagicLens	28.79	<b>34.98</b>	<b>29.31</b>	<b>29.06</b>	18.71	<b>25.63</b>	<b>26.69</b>	<b>34.99</b>
FreeDom	30.73	17.55	28.87	15.55	14.51	12.81	25.49	23.52
BASIC	<b>47.85</b>	<b>39.34</b>	<b>45.79</b>	<b>22.42</b>	<b>30.59</b>	<b>21.99</b>	<b>33.67</b>	<b>38.04</b>

(e) *i*-CIR (visual)

Method	PROJ	DOM	ATTR	APP	VIEW	ADD	CONT
Text	2.09	3.69	2.77	4.13	1.87	4.62	2.75
Image	0.90	0.63	4.37	1.13	6.36	1.53	3.00
Text + Image	4.08	2.46	10.07	3.28	14.30	3.20	19.72
Text $\times$ Image	16.33	23.09	11.53	26.16	30.86	15.56	26.20
Pic2Word	20.02	24.41	11.92	19.02	32.51	16.79	34.55
CompoDiff	3.72	13.49	6.85	15.78	8.53	2.88	16.34
WeiCom	20.55	20.55	13.75	21.88	37.00	9.51	34.56
SEARLE	27.00	16.20	17.20	<b>36.78</b>	36.73	15.69	32.76
CIReVL	27.55	23.66	17.86	29.28	<b>44.37</b>	16.16	27.98
MagicLens	<b>31.05</b>	<b>31.08</b>	<b>24.07</b>	25.76	40.06	<b>28.24</b>	<b>36.40</b>
FreeDom	21.46	22.34	16.42	32.37	17.66	19.04	16.96
BASIC	<b>53.10</b>	<b>39.31</b>	<b>26.29</b>	<b>48.83</b>	<b>47.81</b>	<b>24.04</b>	<b>35.59</b>

(f) *i*-CIR (textual)

#### S4.4 Additional composed image retrieval datasets

**CIRR, FashionIQ, and CIRCO.** Table S5 presents a comparison between BASIC and prior work on CIRR, CIRCO, and FashionIQ. BASIC is outperformed by other approaches, such as CIReVL and CompoDiff, that do not perform well on *i*-CIR. The different datasets, *i.e.*, *i*-CIR, domain conversion datasets, and those in Table S5, have different objectives and reflect different tasks. As a consequence, there is no single approach that is the best among all datasets. We argue that *i*-CIR better reflects real-world applications and use cases.

Unlike *i*-CIR and the domain conversion datasets, the text-only baseline consistently outperforms the image-only baseline. This aligns with our qualitative (Figure S10, Figure S11) and quantitative findings (Figure 6, main manuscript), confirming that in these datasets, the text query often provides sufficient information to resolve the task independently, rather than serving to refine or complement the image query. Notably, in CIRR under the  $R@1$  metric, the text-only baseline even surpasses CompoDiff. Similarly, in FashionIQ, the average performance of the text-only baseline is only 2.65%  $R@10$  below that of FreeDom. CIRCO appears to be less affected by this issue than the other two datasets. To further investigate the imbalance between modalities, we conduct an ablation study by

Table S4: Performance comparison in terms of mAP (%) using SigLIP [18] across five datasets. (a-d): four domain conversion, where for each source domain (columns), we report the average mAP across all corresponding target domains; (e-f): *i*-CIR, where AP performance is averaged over queries grouped by their respective visual (e) or textual (f) category. AVG: average mAP over all source-target domain combinations. **Bold**: best; **purple**: second best.

Method	CAR	ORI	PHO	SCU	TOY	AVG
Text	0.88	0.80	0.62	0.95	0.90	0.83
Image	4.97	3.71	0.85	8.18	7.40	5.02
Text + Image	7.88	5.84	3.08	13.50	12.71	8.60
Text $\times$ Image	6.57	4.34	4.89	6.46	7.46	5.94
FreeDom	49.46	27.12	38.11	47.52	46.90	41.82
BASIC	53.67	39.54	36.40	51.39	53.61	46.92

(a) ImageNet-R

Method	AUT	DIM	GRA	OUT	ROC	WAT	AVG
Text	1.08	1.13	1.04	1.26	1.10	1.11	1.12
Image	6.19	5.19	5.42	7.67	7.44	5.62	6.25
Text + Image	8.35	7.19	8.08	11.42	10.57	8.12	8.95
Text $\times$ Image	2.31	2.91	3.26	3.53	3.25	2.90	3.03
FreeDom	30.28	29.96	33.86	37.16	33.14	26.49	31.81
BASIC	28.11	32.39	31.13	34.07	29.26	23.12	29.68

(c) NICO++

Method	FICT	LAND	MOB	HOUSE	TECH	FASH	PROD	ART
Text	7.61	5.09	2.18	11.59	4.40	8.33	1.82	1.59
Image	1.31	3.46	2.91	6.32	2.42	5.96	2.03	1.52
Text + Image	3.91	13.41	6.92	12.51	12.90	11.71	7.89	3.58
Text $\times$ Image	25.56	21.42	12.93	32.86	28.53	18.30	12.18	22.67
FreeDom	27.75	27.10	19.32	43.02	31.36	35.31	20.27	33.11
BASIC	50.65	53.45	39.56	48.87	50.84	52.83	45.11	44.43

(e) *i*-CIR (visual)

Method	CLIP	PAINT	PHO	SKE	AVG
Text	0.76	0.72	0.76	0.75	0.74
Image	5.07	7.53	3.68	6.15	5.61
Text + Image	7.79	11.33	10.80	9.02	9.74
Text $\times$ Image	3.00	2.60	4.34	3.18	3.28
FreeDom	57.14	45.47	59.71	52.21	53.63
BASIC	59.76	45.58	54.67	47.73	51.94

(b) MiniDN

Method	TODAY	ARCHIVE	AVG
Text	5.02	3.84	4.43
Image	28.14	10.25	19.20
Text + Image	26.73	10.16	18.44
Text $\times$ Image	3.49	4.87	4.18
FreeDom	47.00	27.45	37.22
BASIC	45.01	39.09	42.05

(d) LTLL

Method	PROJ	DOM	ATTR	APP	VIEW	ADD	CONT
Text	5.95	5.14	5.22	6.88	1.74	10.45	4.55
Image	1.24	1.09	6.31	1.35	8.98	4.09	4.63
Text + Image	7.37	3.52	11.18	3.29	17.12	12.66	20.12
Text $\times$ Image	21.61	17.89	20.19	26.47	11.15	28.24	25.13
FreeDom	22.94	25.25	20.03	31.52	38.34	39.17	25.59
BASIC	53.41	51.17	42.19	50.62	63.73	47.42	50.90

(f) *i*-CIR (textual)

removing the Harris regularization (denoted as BASIC \*). Results show improved performance in CIR and FashionIQ, supporting the hypothesis of modality imbalance.

**Refined CIR and FashionIQ.** We evaluate BASIC and BASIC without Harris regularization (BASIC \*) on the newly released CoLLM [4] refined [4] versions of CIR and FashionIQ in Table S6. Both variants gain notably on the refined splits (*e.g.*, up to +13.7  $R@10$  on refined-FashionIQ SHIRT). These gains are consistent with CoLLM’s goal of reducing annotation ambiguity. This is orthogonal to modality dominance (as evidenced by BASIC \* outperforming BASIC on text-dominant settings) and to CLIP’s phrasing sensitivities (*e.g.*, instructional/relational text).

#### S4.5 Dedicated vs generic corpora

We further explore the impact of using dataset or domain-specific (dedicated) vs. generic corpora on retrieval performance. This analysis is conducted on LTLL, which contains a narrow range of object categories—specifically, landmarks such as monuments and architectural sites.

The goal is to assess whether adding more targeted corpora can improve performance. While incorporating a landmark-only corpus into *i*-CIR would be problematic, due to the large number of irrelevant distractors that can introduce noise when projecting into a restricted semantic space, LTLL is composed entirely of landmark-related imagery, making it well-suited for such decomposition.

We experiment with two types of corpora:

- *Generic corpora*, which include general stylistic descriptors.
- *Dedicated corpora*, designed specifically for this task:

Table S5: Performance comparison using CLIP [10] across three generic composed image retrieval datasets. BASIC is evaluated for the default settings, as well as without the Harris regularization (\*), which found to be beneficial for these datasets.

Method	R@1	R@5	R@10	R@50	Method	mAP@5	mAP@10	mAP@25	mAP@50
Text	20.96	44.89	56.80	79.16	Text	3.09	3.25	3.76	4.01
Image	7.42	23.61	34.07	57.40	Image	1.60	2.02	2.76	3.13
Text + Image	12.41	36.15	49.18	78.27	Text + Image	4.06	5.20	6.29	6.85
Text $\times$ Image	22.55	50.36	62.84	86.02	Text $\times$ Image	11.64	12.29	13.64	14.28
Pic2Word	23.90	51.70	65.30	87.80	Pic2Word	8.70	9.50	10.70	11.30
SEARLE	24.20	52.50	66.30	88.80	SEARLE	11.70	12.70	14.30	15.10
CompoDiff	18.20	53.10	70.80	90.30	CompoDiff	12.60	13.40	15.80	16.40
FreeDom	21.00	48.70	61.90	88.10	FreeDom	14.00	14.80	16.40	17.20
CIReVL	24.60	52.30	64.90	86.30	CIReVL	18.60	19.00	20.90	21.80
MagicLens	30.10	61.70	74.40	92.60	MagicLens	29.60	30.80	33.40	34.40
BASIC	15.83	40.89	53.90	82.27	BASIC	15.95	16.77	18.19	18.94
BASIC *	17.98	44.92	58.80	86.51	BASIC *	15.95	16.77	18.21	19.00

(a) CIRR

(b) CIRCO

Method	Dress		Shirt		Toptee		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Text	14.53	32.92	20.51	34.69	21.83	39.57	18.95	35.73
Image	4.36	12.84	10.55	19.97	8.21	16.37	7.71	16.39
Text + Image	17.40	35.60	21.84	36.26	23.30	39.37	20.85	37.08
Text $\times$ Image	21.52	41.00	27.63	42.05	28.71	47.22	25.95	43.42
Pic2Word	20.00	40.20	26.20	43.60	27.90	47.40	24.70	43.70
SEARLE	20.50	43.10	26.90	45.60	29.30	50.00	25.60	46.20
CompoDiff	32.20	46.30	37.70	49.10	38.10	50.60	36.00	48.60
FreeDom	16.80	36.30	23.50	38.50	24.70	43.70	21.60	39.50
CIReVL	24.80	44.80	29.50	47.40	31.40	53.70	28.60	48.60
MagicLens	25.50	46.10	32.70	53.80	34.00	57.70	30.70	52.50
BASIC	18.59	37.68	26.30	44.80	23.92	40.95	22.94	41.14
BASIC *	19.98	39.86	29.88	47.06	26.21	44.57	25.36	43.83

(c) FASHIONIQ

Table S6: Performance on CoLLM-refined CIRR and FashionIQ. Both BASIC and BASIC \* improve on the refined splits, indicating reduced annotation ambiguity.

Dataset	Split	Method	mAP	R@1	R@5	R@10	R@50
CIRR	Legacy	BASIC	24.34	15.83	40.89	53.90	82.27
	Refined	BASIC	28.28	21.87	48.96	60.82	86.66
	Legacy	BASIC *	26.89	17.98	44.92	58.80	86.51
	Refined	BASIC *	31.89	25.30	53.49	65.94	90.15
FashionIQ DRESS	Legacy	BASIC	6.96	–	–	18.59	37.68
	Refined	BASIC	9.91	–	–	24.40	47.63
	Legacy	BASIC *	7.43	–	–	19.98	39.86
	Refined	BASIC *	11.27	–	–	28.59	52.43
FashionIQ SHIRT	Legacy	BASIC	10.93	–	–	26.30	44.80
	Refined	BASIC	17.15	–	–	36.82	56.09
	Legacy	BASIC *	12.93	–	–	29.88	47.06
	Refined	BASIC *	22.55	–	–	43.54	62.40
FashionIQ TOPTEE	Legacy	BASIC	10.28	–	–	23.92	40.95
	Refined	BASIC	14.57	–	–	32.08	51.41
	Legacy	BASIC *	12.08	–	–	26.21	44.57
	Refined	BASIC *	19.08	–	–	38.27	57.98

- Architectural terms (e.g., “Byzantine basilica”, “Crusader fortress”, “slate tile”, “oxidized copper”, “weathered limestone”)
- Temporal/stylistic cues (e.g., “flash photography”, “shadows of age”, “digital camera photo”, “restored image”)

Table S7: LTLL dataset mAP (%) results using different combinations of positive (+) and negative (−) text corpora: generic (Gen) or dedicated (Ded). We consider different text queries with the same semantic meaning. Rows correspond to target domains - which are the text queries used.

Target Domain	Only Gen+	Gen+ & Gen−	Only Ded+	Ded+ & Ded−	Gen+ & Ded−	Ded+ & Gen−
archive	41.41	40.00	42.83	43.00	43.60	40.97
today	42.13	42.76	43.96	45.97	44.38	45.74
<i>mean</i>	41.77	41.38	43.40	44.48	43.99	43.36
old	50.52	48.86	50.78	50.77	50.58	48.89
new	38.24	40.00	41.77	45.68	41.12	44.66
<i>mean</i>	44.38	44.43	46.28	48.22	45.85	46.78
vintage	54.04	53.92	54.30	57.37	57.43	53.86
today	42.13	42.76	43.96	45.97	44.38	45.74
<i>mean</i>	48.08	48.34	49.13	51.67	50.90	49.80

We also evaluate the effect of varying textual queries that convey the same underlying meaning, such as synonyms or stylistic variants. Examples include:

- “archive” vs. “old” vs. “vintage” — all describing temporally distant or aged imagery.
- “today” vs. “new” — both indicating contemporary or modern representations.

Table S7 presents these performance comparisons over different combinations of corpora and text queries. The following observations can be made:

- Different textual queries, despite their close-almost identical-semantic meaning, can have a substantial impact on retrieval results. The results for the first two textual pairs (“today” vs. “now” and “archive” vs. “old”) reveal interesting trends. The query “today” is more effective at retrieving contemporary photos compared to “now”. In contrast, the “archive” underperforms relative to “old”. Semantically, this is expected—“archive” is not commonly used to describe old photos in natural language queries. To address this, we also include the query “vintage”, which offers a clearer and more intuitive descriptor for aged or historical imagery.
- The use of dedicated corpora consistently improves performance across all textual variants. Even a single dedicated corpus yields a notable boost.
- Adding a negative corpus generally enhances performance, with the exception of the (“today”, “archive”) pair when combined with the generic stylistic corpus.
- The baseline mAP for the (“today”, “archive”) setting is 41.38% using generic corpora. With dedicated corpora and well-matched textual queries the performance increases to 51.67%, representing a gain of more than 10% mAP.

We further examine the case of *i*-CIR by leveraging its predefined visual categories, as illustrated in Figure S1. Based on this taxonomy, we construct a dedicated object corpus of 300 elements tailored specifically to *i*-CIR. This setup serves as a proof-of-concept, as it assumes access to detailed visual categorization—information that is typically unavailable in realistic retrieval scenarios.

To avoid introducing bias toward any specific category, which may collapse the semantic representation towards a specific direction, we ensure that objects are selected from across all visual categories. As in previous experiments, we concatenate the dedicated corpus entries with those from the generic corpus.

Table S8: *i*-CIR performance comparison of different object corpora in terms of mAP (%). CLIP [10] backbone is used. We compare the case of the generic corpus vs the dedicated corpus across the different visual categories of *i*-CIR.

Positive Corpus	FICT	LAND	MOB	HOUSE	TECH	FASH	PROD	ART
Generic	47.85	39.34	45.79	<b>22.42</b>	30.59	21.99	33.67	<b>38.04</b>
Dedicated	<b>49.48</b>	<b>39.35</b>	<b>47.05</b>	21.92	<b>34.45</b>	<b>22.91</b>	<b>35.00</b>	37.88

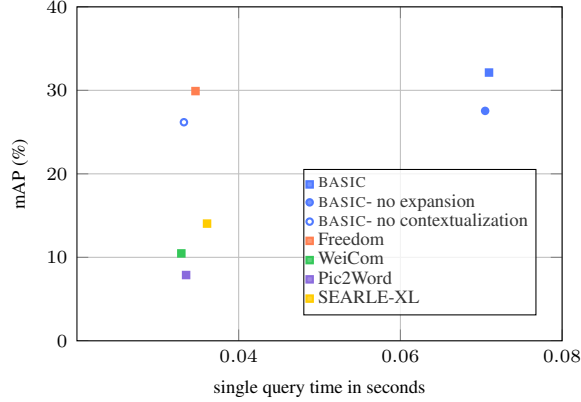


Figure S5: *Latency vs. mAP(%) on ImageNet-R. Comparison of methods from the literature and BASIC with and without query expansion and contextualization.*

**Table S8** reports per-category mAP on *i*-CIR. The category-aware dedicated corpus yields improvements in six of eight categories (parity in LANDMARK, minor drops in HOUSEHOLD and ART). Note that these corpora were automatically generated with ChatGPT from a short prompt (see example in box) and were not hand-balanced to match *i*-CIR’s taxonomy; their distribution only loosely aligns with the actual category frequencies. Despite this mismatch—and the fact that we simply concatenate the dedicated with the generic corpus—we still observe consistent gains, suggesting that a weak, category-aware prior can help without tightly fitting the dataset. We view this as a proof-of-concept.

#### Dedicated word corpus prompt

Provide a .txt file with 300 distinct household objects (as if they are labels for a classification task). Be broad and diverse.

### S4.6 Time requirements

**Figure S5** presents a per-query latency comparison between BASIC and several existing methods from the literature: WeiCom, Pic2Word, SEARLE-XL, and Freedom, evaluated on ImageNet-R. BASIC achieves the highest performance, reaching a mAP of 32.13 with a latency of 70.96ms per query. When query expansion is disabled, the performance drops by 4.59 mAP, while the latency improves slightly by 0.48ms. The primary overhead in BASIC is the contextualization step. When removed, BASIC (33.23ms) becomes faster than Freedom (34.66ms), SEARLE-XL (36.09ms), and Pic2Word (33.50ms), and only marginally slower than WeiCom by 0.03ms. Note that no further retrieval optimization has been applied for these experiments (*e.g.* FAISS [5]).

We intentionally exclude CompoDiff [3] and CIREVL [6] from this latency comparison: CompoDiff relies on a diffusion generator at inference, and CIREVL invokes a captioner (BLIP 2 [7]) and an LLM (LLaMA [14]). These pipelines are orders of magnitude heavier and are not designed for low-latency retrieval, so including them would make the plot incomparable and obscure the efficiency trade-offs among lightweight methods.

As expected, aside from the contextualization step — which introduces the overhead of computing additional CLIP embeddings (fixed to generating 100 such descriptions across all experiments) — the proposed approach remains very lightweight. The overhead introduced by contextualization is a current limitation of the method, but could be addressed by integrating a separate shallow network that contextualizes the CLIP embedding of a given text query, as discussed in [section S3](#).

### S4.7 *i*-CIR: Compact but hard.

We use randomly selected images from LAION as negatives to assess how challenging *i*-CIR is in comparison to a large-scale database that is commonly shared across all queries and lacks explicit hard negatives. **Figure S6** shows that the performance drop on *i*-CIR is comparable to the degradation

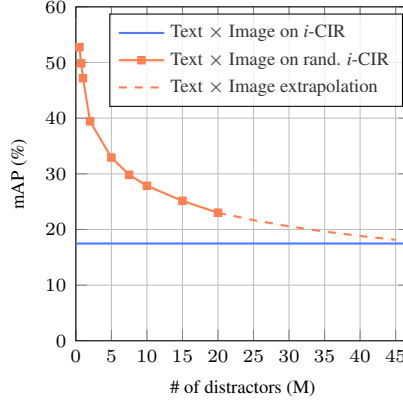


Figure S6: *Impact of curated vs. random negatives on retrieval difficulty.* We plot mAP on *i*-CIR with its 750K curated hard negatives (blue) against the same benchmark augmented with up to 20M random LAION distractors (orange, solid), with a smooth regression trend shown dashed, reaching more than 40M. Even tens of millions of distractors do not match the difficulty imposed by curated hard negatives.

caused by tens of millions of randomly sampled LAION images. Using the Text  $\times$  Image baseline, we show that *i*-CIR, with its curated hard negatives, is substantially more challenging than a 20M-image LAION distractor set. A regression analysis further indicates that *i*-CIR has difficulty comparable to using approximately 45M randomly sampled negatives.

#### S4.8 Collection bias.

While we take explicit steps to mitigate potential bias toward CLIP ViT-L/14-based methods during data collection such as discarding all seed images and ensuring that seed sentences do not exactly match any text query, the possibility of residual bias cannot be entirely ruled out. Since the retrieval of candidate negatives relies on CLIP ViT-L/14, one might expect this model to underperform relative to weaker models due to the increased difficulty of the curated negatives tailored to its embedding space. To test this, we evaluate the standard Text  $\times$  Image baseline using a weaker model, CLIP ViT-B/32. The results show that ViT-B/32 achieves 18.66% mAP on *i*-CIR, notably lower than the 23.28% mAP of ViT-L/14. This confirms that no bias favoring alternative models has been introduced and that the performance of ViT-L/14 is not artificially deflated by the collection process.

#### S4.9 Effect of Text Query Style on Retrieval

A text query can be almost equivalently formulated in different ways, *e.g.*, concisely (“engraving”), in a longer sentence (“as a vintage stylish engraving”), in a relational way (“the same landmark as the one depicted in the image but as an engraving”), and in an instructional way (“engrave the landmark depicted in the image”). In the context of *i*-CIR, we refer to those cases as *original*, *longer*, *relational*, and *instructional*, respectively. The queries of *i*-CIR are brief and comprehensive, while existing datasets like CIRR and CIRCO are often relational or instructional, while also having larger lengths.

To probe the robustness of BASIC against such text query styles, we automatically rewrite (via ChatGPT) all text queries of *i*-CIR<sub>named</sub>, and we present the results in Table S9 ( $\Delta$  is the relative change with respect to the original row).

**Longer.** We expand all text queries by adding 1 to 4 extra words while preserving their meaning. On average the phrases grow from 4.4 to 7.9 words, an increase of 3.5 words or +79%, effectively almost doubling their length. The median rises from 4 to 8 words, the shortest query goes from 1 to 4 words, and the longest from 13 to 15. Example: “crowded” becomes “in a dense and crowded scene”. The ChatGPT prompt used to convert text queries is presented below.



Table S9: *Impact of text query style on retrieval.* We compare the original (concise) queries from *i-CIR*<sub>named</sub> with automatically rewritten *longer*, *relational*, and *instructional* variants and report mAP for BASIC, FreeDom, and MagicLens.  $\Delta$  is the relative change vs. the original row.

Text-query Variant	BASIC mAP	BASIC $\Delta\%$	FreeDom mAP	FreeDom $\Delta\%$	MagicLens mAP	MagicLens $\Delta\%$
Original (concise)	47.39	—	25.14	—	32.00	—
Longer	45.02	−5	23.91	−5	26.90	−16
Relational	38.95	−18	18.99	−25	29.34	−8
Instructional	37.47	−21	18.31	−27	28.85	−10

#### Longer text query prompt

Take each text query and, without altering its core meaning or introducing new concepts, enrich it with tasteful descriptive words or clarifying phrases so that its length is roughly doubled. Preserve the original order and casing and return each pair as “original query”: “augmented longer query” with no additional commentary. Example: “crowded” → “in a dense and crowded scene”

We observe that BASIC has a small performance drop similar to that of FreeDom, while MagicLens has a larger drop, therefore revealing a *new advantage* of training-free methods (relying on CLIP). Related to the above question, MagicLens presumably has a larger drop due to shorter sentences in its training.

**Relational.** To give every text query an explicit link to its paired reference image, we rewrite all text queries by prepending a short relational clause while keeping the original order and casing intact. Example: “with fog” becomes “the same object instance depicted in the image but with fog”. This yields a suite of explicitly relational captions that stress compositional understanding without altering the original intent. The ChatGPT prompt used to convert text queries is presented below.

#### Relational text query prompt

Rewrite every text query so that it explicitly refers to the same object instance of the image query, while preserving original words in the same order. Simply prepend a concise relational clause, avoid introducing new concepts or proper names, and return each pair as “original query”: “relational query”, with no additional commentary. Example: “with fog” becomes “the same object instance depicted in the image but with fog”

We observe that the drop of BASIC is noticeably smaller than that of FreeDom, but also greater than that of MagicLens. This pronounces the benefit of training for composed image retrieval, and a drawback of training-free methods, which is due to the fact that there is no or little relational information during the CLIP pre-training. Nevertheless, BASIC still performs reasonably well.

**Instructional.** We recast every text query as an imperative instruction that tells the system what to do with the reference-image object. Concretely, we prepend an action verb (*e.g.*, “turn”, “place”, “render”) and then a clause like “the object depicted in the image”, then append the original words in the original order. Example: “in a live action scene” becomes “adapt the object depicted in the image into a live action scene”. The ChatGPT prompt used to convert text queries is presented below.

#### Instructional text query prompt

Rewrite each text query as a concise imperative instruction that starts with an appropriate action verb, explicitly mentions something like “the object depicted in the image”, preserves the original



words in the same order, introduces no new concepts, and output each pair exactly as ‘original query’: ‘instructional query’ with no commentary.

The performance pattern across the methods in this case mirrors the relational case.

Additionally, we perform the reverse mapping for 10 manually selected relational queries of CIRRE that are converted into shorter and absolute descriptions. For example, ‘Make the dog older and have two birds next to him and make everything look like a painting’ becomes ‘as a painting of an older dog with two birds’ and ‘Change the background to trees and remove all but one dog sitting in grass looking right’ becomes ‘as a single dog sitting in grass with trees in the background’.

Interestingly, evaluating using the 10 rewritten queries of CIRRE (shown in Table S10) to make them non-relational (concise) demonstrates a large performance increase for BASIC from 18% mAP (10% R@1) to 44.4% mAP (40% R@1). This signifies that the relational description might not be the best way to describe those queries either.

Table S10: *Making relational CIRRE queries concise*. Converting 10 relational CIRRE queries into concise, non-relational descriptions substantially improves BASIC, shown via mAP and recall.

Text-query Variant	mAP	R@1	R@5	R@10	R@50
Relational	18.0	10.0	40.0	60.0	80.0
Concise	44.4	40.0	60.0	70.0	90.0

#### S4.10 Retrieval visualizations across BASIC components.

Figure S7 and Figure S8 qualitatively demonstrate the effects of centering and the performance of the full BASIC compared to the Text  $\times$  Image baseline on *i*-CIR. Notably, even centering alone yields substantial performance gains.

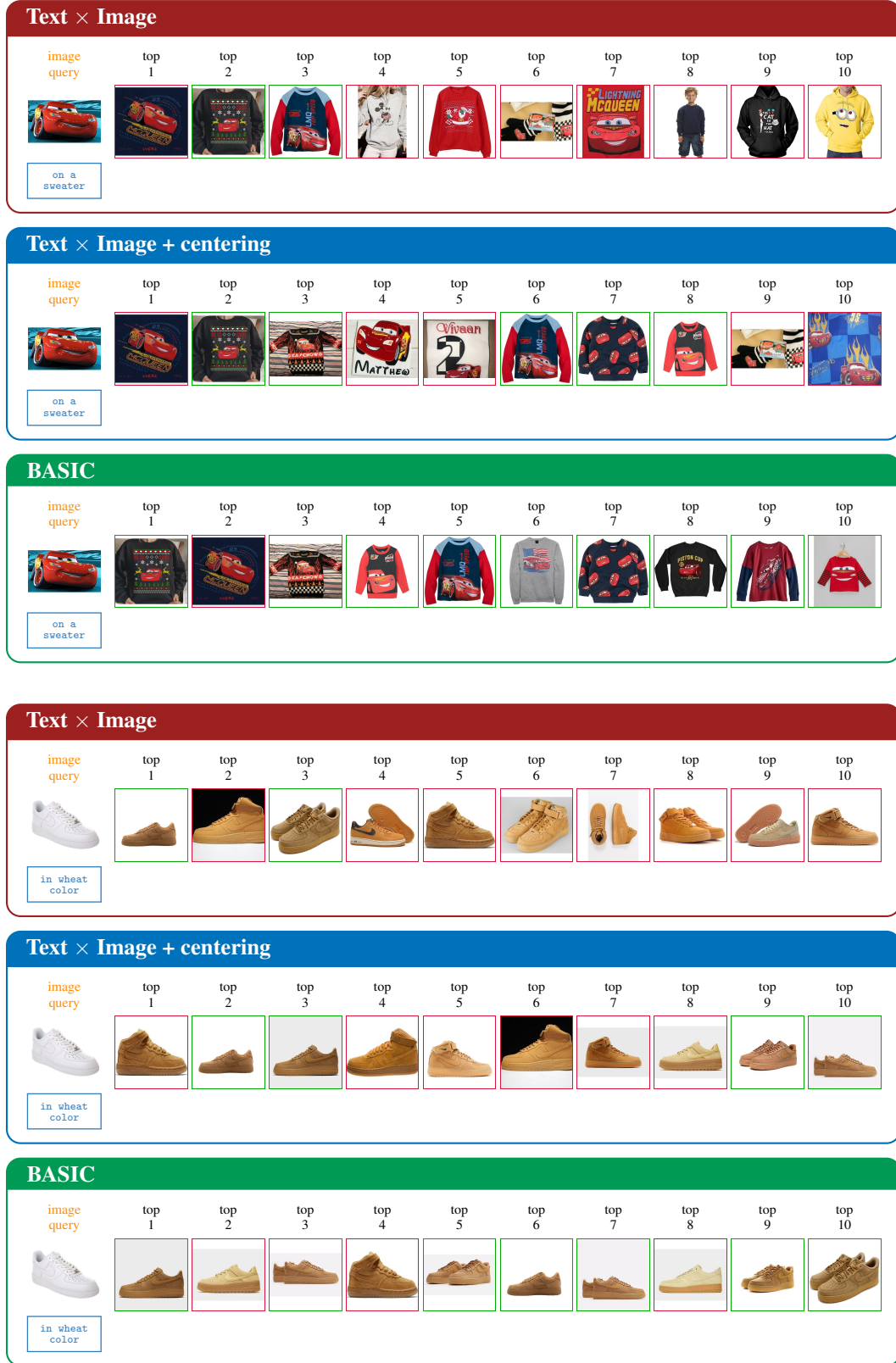


Figure S7: Retrieval results for Text × Image, Text × Image with centering, and BASIC on *i*-CIR.

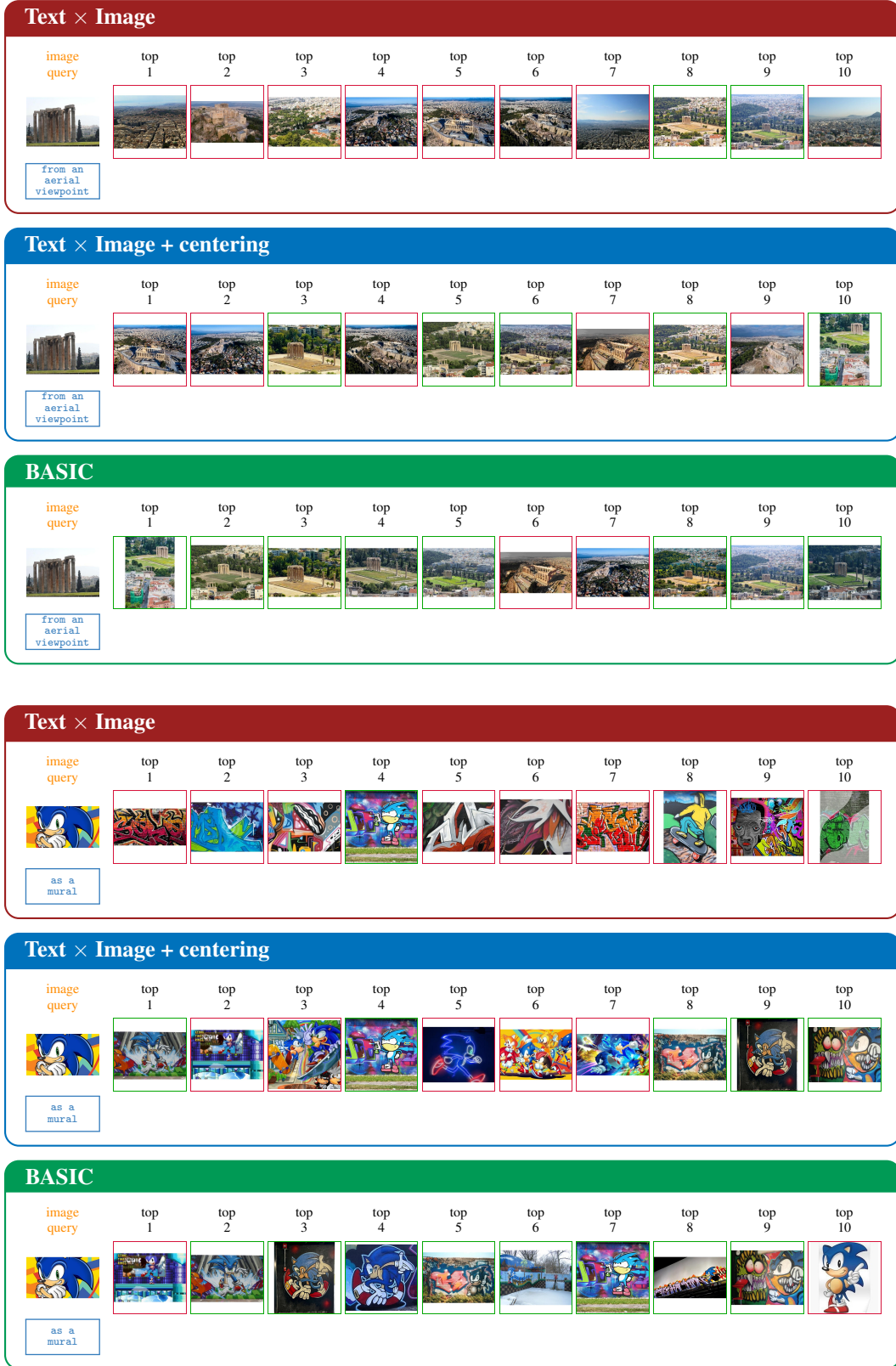


Figure S8: Retrieval results for Text  $\times$  Image, Text  $\times$  Image with centering, and BASIC on *i*-CIR.

- along with the tetris game cartridge
- arranged beside the bed
- as a balloon
- as a billboard advertisement
- as a bobblehead
- as a bobblehead toy
- as a colored pencil drawing
- as a colorful artistic illustration
- as a colorful digital illustration
- as a colorful sketch
- as a colorized artistic interpretation
- as a decorative piece in bowl
- as a decorative piece on a table
- as a detailed line drawing
- as a digital 2D illustration
- as a digital drawing
- as a digital illustration
- as a digital illustration with a Christmas tree in front
- as a digital photomosaic
- as a digital watercolor painting
- as a figurine
- as a flat vector orthographic digital illustration
- as a freehand monochromatic sketch
- as a giant balloon
- as a graffiti mural
- as a historical monochromatic drawing
- as a huge balloon
- as a huge floating balloon
- as a human size sculpture
- as a human size statue
- as a lego
- as a lego brick figure
- as a lego painting
- as a live action adaptation
- as a mascot costume
- as a miniature
- as a monochromatic sketch
- as a monochrome engraving
- as a monochrome sketch
- as a mural
- as a napkin holder
- as a painting
- as a painting including its surroundings
- as a paper folding art
- as a pen sketch
- as a pencil drawing
- as a photo during a Christmas night
- as a photo during day
- as a photo reflecting on water
- as a photorealistic 3d model wearing a hat
- as a pixel art graffiti mural
- as a plushie
- as a pop art poster
- as a professional mascot costume
- as a professional walk-around mascot costume
- as a public art installation
- as a real size statue
- as a rendered 3D model
- as a sand sculpture
- as a scale model
- as a sculptured edible figure on a cake
- as a silhouette with reflection on water
- as a sketch
- as a street art mural
- as a stylized digital illustration in a framed print
- as a stylized digital-paint illustration
- as a technical schematic design
- as a toy
- as a travel poster art
- as a vector clipart-style digital illustration
- as a vintage postcard
- as a wash drawing or an aquatint
- as a watercolor painting
- as a watercolor style illustration
- as an AI generated digital illustration
- as an action figure
- as an architectural plan
- as an archive photo
- as an emoji
- as an engraving
- as an iconic rooftop sign
- as an oil painting
- as an old archival photo
- as digital illustration
- as japanese art
- as large scale outdoor installation
- as oil painting
- as pencil drawing
- as pixel art
- as seen from a city road
- as vivid pop art
- at day
- at day with blue sky
- at night
- at night with a sculpture in front
- at night with full moon rising behind
- at night with the clock showing around 8:40
- at sunset
- at the evening reflecting on the water
- attached to keys
- being deflated
- being lit
- being wet
- being worn
- broken
- carried by a person
- clipped on a striped bag
- closed
- collapsed and leaned against a wall
- compressed into a flat form
- containing an old photograph
- covered with a black hat
- covered with a decorative embroidered cloth
- covered with a towel
- covered with soil in a garden
- covered with stones
- crowded
- decorated with flags
- disassembled
- displayed on a nightstand
- during day
- during day with blue sky
- during liftoff
- during night
- during sunset
- during the day reflecting on the water
- during the making
- filled with clothes
- filled with coffee
- filled with cosmetics
- filled with dog food
- filled with food leftovers
- filled with orange juice
- filled with sugar
- filled with water
- floating on seawater
- floating on the surface of seawater
- folded
- folded with other clothes
- from a 3D animation movie
- from a drone shot
- from a fully vertical top-down perspective
- from a low-angle perspective
- from a top-down fully vertical perspective
- from an aerial perspective
- from an aerial viewpoint
- from an aerial viewpoint during day
- hanging from a flower bouquet
- hanging from a tree branch
- hanging on a clothesline
- hanging on a clothesline outdoors
- hanging on a drying rack
- hanging on a folding beach chair
- hanging on a t-shirt
- hanging on a wall
- hanging on fireplace guard
- hanging on the drying rack
- hanging on the key holder
- hanging on the wall
- having mud and dirt marks
- held in hand
- held in hand on a beach
- held in hands
- holding a bunch of bananas
- holding a chicken
- holding bars of soap
- holding hair
- holding two espresso cups
- holding various clothes
- illuminated
- in a cabrio version
- in a closed position
- in a comic panel
- in a deep state of rest
- in a desk organizer
- in a filled state
- in a folded state
- in a fully unfolded position
- in a garden pot
- in a laundry tub
- in a manga panel
- in a person's ear
- in a vintage magazine advertisement
- in a vintage magazine page
- in a wet state
- in an old archival photo
- in an open state
- in archive photo
- in black and white color with gum outsole
- in black color
- in front of buildings
- in golden color
- in golden yellow color
- in green color
- in pink color
- in purple color
- in solid yellow color
- in the original video game
- in turquoise color on a black t-shirt
- in wheat color
- in white color
- inside a case
- inside a dishwasher
- inside a plastic bag
- inside a small bag
- inside a wallet
- inside a washing machine
- inside an open suitcase
- inside its original packaging
- inside their case
- lit up with nighttime projection mapping visuals
- lying open
- next to a dog
- next to a plant
- next to coffee beans
- next to other bottles
- next to other fishing boats
- next to other kitchen utensils
- next to other sunglasses
- next to other toys
- next to plants
- next to the real animal it represents
- on a Christmas sweater
- on a backpack
- on a billboard ad
- on a black t-shirt
- on a blue t-shirt
- on a bookshelf
- on a curtain
- on a desk next to a laptop
- on a fireplace
- on a flyer
- on a hat
- on a mug
- on a sweater
- on a t-shirt
- on a t-shirt with a cartoonish style
- on a table next to food
- on a white covered armchair
- on a white t-shirt
- on a yellow t-shirt
- on clothing
- on hanger
- on watermelon shaped dog bed
- outdoors on grass
- packed into its cardboard box
- partially flipped upside down
- placed around a table
- placed in a garden
- placed in a shelf
- placed inside a glass jar
- placed on a metal chair
- placed on a metal tray
- placed on a plastic chair
- placed on a wall-mounted holder
- placed on top of clothes
- plugged in and glowing light
- positioned in front of an office desk
- printed on a white pillow
- reflected in a mirror
- reflecting on water
- riding a white dinosaur
- rolled-up in a bag
- sitting on a chair
- smashed
- spread on the bed
- spread out on a rocky ground
- stacked on shelves
- stained with coffee
- stored in a nightstand
- stored in a wardrobe
- surrounded by chairs
- surrounded by metallic kitchenware
- the mountain depicted on the packaging
- tied around a ponytail
- unassembled
- under running water
- used as a seat cushion
- while charging
- while flying
- with a candle on top
- with a close up on the bridge
- with a crowd in front
- with a darker mask and a hoodie
- with a dog inside
- with a film around
- with a glass on top
- with a hair claw clip clipped onto it
- with a lot of people around
- with a man holding it
- with a man proposing to a woman in front
- with a more rounded handle
- with a numeric keypad
- with a patio umbrella open
- with a person pretending to hold it up
- with a person sitting on
- with a real car in front
- with a seaside background
- with a towel draped over
- with an external flash attached
- with bananas placed on it
- with big waves crashing against the shoreline
- with blond fur
- with brown or dark beige leather
- with candies on top
- with doors open
- with drawers open
- with fireworks
- with fog
- with people around
- with people in front
- with people standing on
- with seat cushion
- with snow
- with the front wheel removed
- with the tongue out
- with various ornaments displayed on top
- with waves crashing against it
- with yellow shoelaces
- without cushion
- without seat cushion and pillows
- without sticky paper notes but filled with pens and markers
- without sunset
- worn
- worn around a dog's neck
- worn around a human's neck
- worn by a child
- worn by a dog
- worn by a girl
- worn by a person
- worn by someone
- worn on a face sculpture
- worn on the ankle
- worn with colorful swimwear

Figure S9: *The unique text queries of i-CIR dataset*, listed in alphabetical order. These queries reflect diverse compositional transformations across appearance, context, attributes, viewpoint, and more.

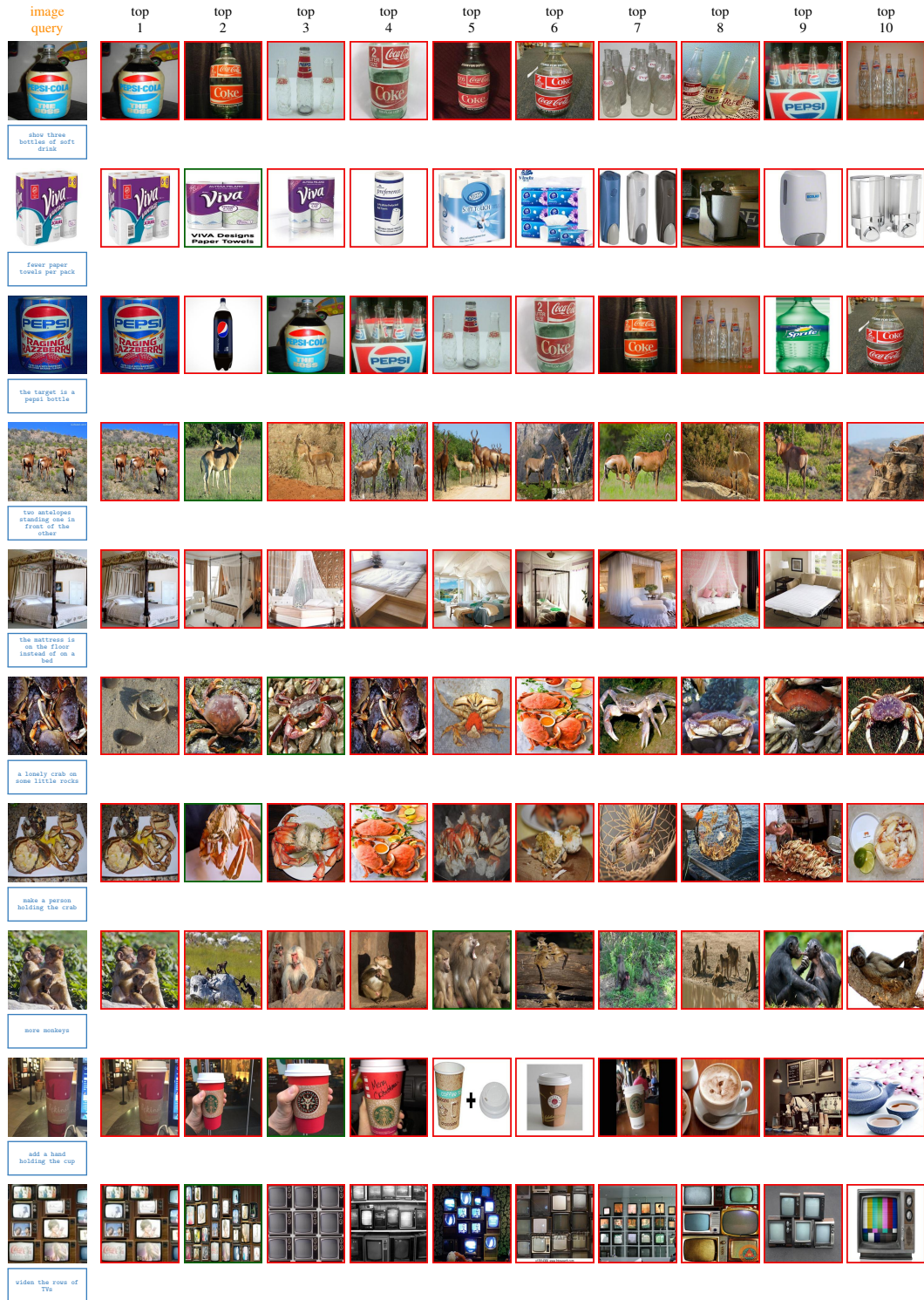


Figure S10: CIRR retrieval results using  $\text{Text} \times \text{Image}$ , illustrating dataset limitations. Each row shows a composed query with its top-10 retrieved results. **Green** boxes indicate correct retrievals, **red** denote incorrect ones. Several cases reveal issues with ground-truth annotations and the limited role of the image query.





Figure S11: *FashionIQ* retrieval results using  $\text{Text} \times \text{Image}$ , highlighting dataset limitations. Each row shows a composed query with image (left) and text (side). The top-10 retrieved results are shown left to right. Green boxes mark correct retrievals, red boxes mark incorrect ones. The figure highlights false negatives and limited reliance on the image query.

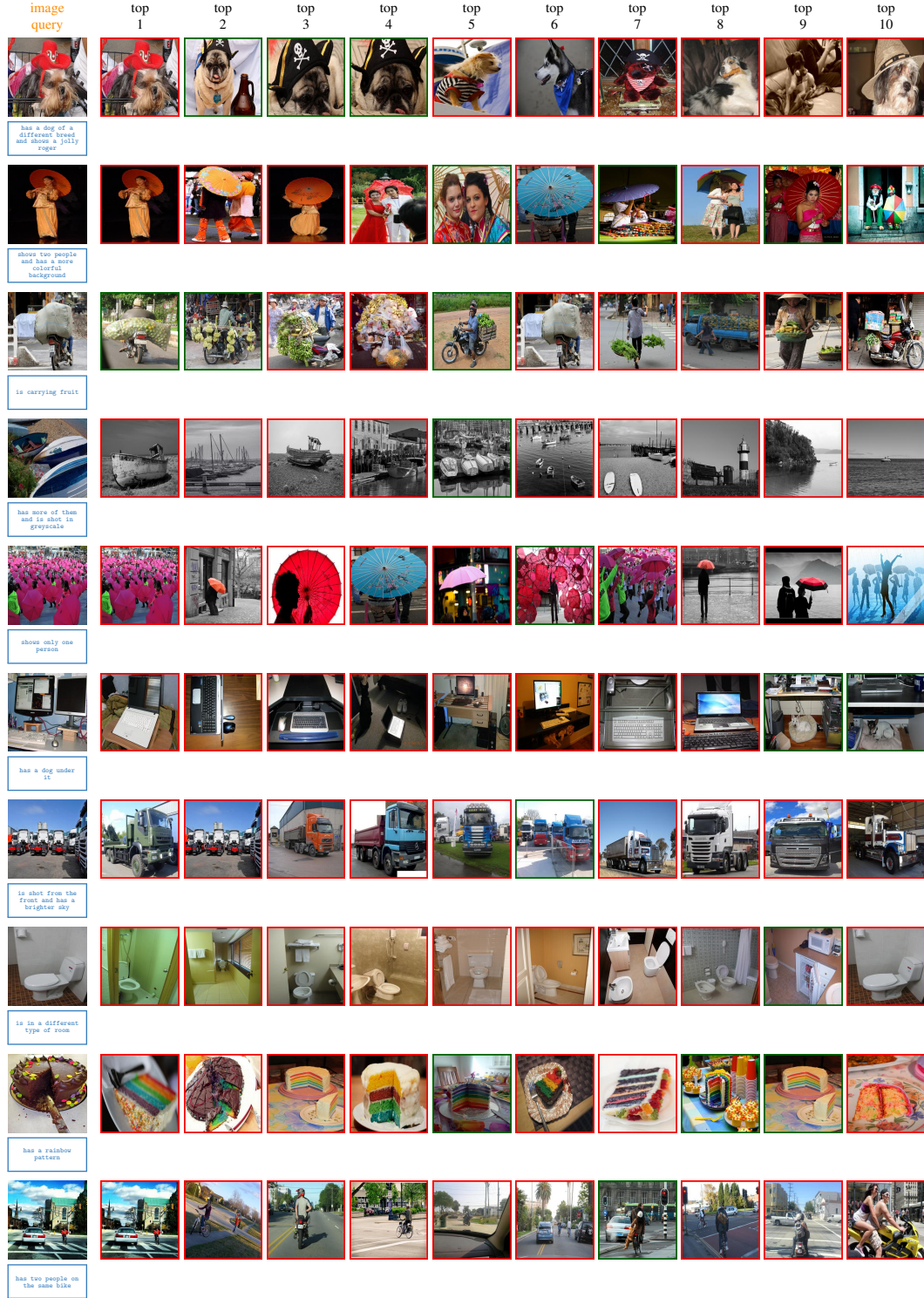


Figure S12: *CIRCO* retrieval results using  $\text{Text} \times \text{Image}$ , highlighting dataset limitations. Each row shows a composed query with image (left) and text (side). The top-10 retrieved results are shown left to right. Green boxes mark correct retrievals, red boxes mark incorrect ones. The figure highlights false negatives and the instructional nature of the text query.



## References

- [1] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, 2023.
- [2] Mikolaj Czerkawski and Alistair Francis. From laion-5b to laion-eo: Filtering billions of images using anchor datasets for satellite image extraction. In *ICCV Workshop.*, 2023.
- [3] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023.
- [4] Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. Collm: A large language model for composed image retrieval. In *CVPR*, 2025.
- [5] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. In *arXiv*, 2017.
- [6] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *arXiv*, 2023.
- [7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.
- [8] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, 2021.
- [9] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.
- [10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.
- [11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
- [12] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022.
- [13] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *arXiv*, 2015.
- [14] 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.
- [15] Shuang Wang and Shuqiang Jiang. Instre: a new benchmark for instance-level object retrieval and recognition. *ACM TOMM*, 11:37, 2015.
- [16] Tobias Weyand, André Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 - A large-scale benchmark for instance-level recognition and retrieval. In *CVPR*, 2020.
- [17] 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 *CVPR*, 2021.
- [18] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.