Instance-Level Composed Image Retrieval

Bill Psomas[1*] **George Retsinas**[2*] **Nikos Efthymiadis**[1] **Panagiotis Filntisis**[2,4]
Yannis Avrithis[5] **Petros Maragos**[2,3,4] **Ondrej Chum**[1] **Giorgos Tolias**[1]

[1]VRG, FEE, Czech Technical University in Prague [2]Robotics Institute, Athena Research Center
[3]National Technical University of Athens [4]Hellenic Robotics Center of Excellence [5]IARAI

Figure 1: We introduce *i*-nstance-level Composed Image Retrieval (*i*-CIR) evaluation dataset. Given an image query depicting a specific instance (*e.g.*, Temple of Poseidon) along with a modifying text query, the task is to retrieve images showing the same instance altered according to the text (composed positives). Unlike existing datasets [39, 25, 2], *i*-CIR explicitly ensures the presence of challenging negative examples across three distinct dimensions: visual, textual, and composed.

Abstract

The progress of *composed image retrieval* (CIR), a popular research direction in image retrieval, where a combined visual and textual query is used, is held back by the absence of high-quality training and evaluation data. We introduce a new evaluation dataset, *i*-CIR, which, unlike existing datasets, focuses on an instance-level class definition. The goal is to retrieve images that contain the same particular object as the visual query, presented under a variety of modifications defined by textual queries. Its design and curation process keep the dataset compact to facilitate future research, while maintaining its challenge—comparable to retrieval among more than 40M random distractors—through a semi-automated selection of hard negatives. To overcome the challenge of obtaining clean, diverse, and suitable training data, we leverage pre-trained vision-and-language models (VLMs) in a training-free approach called BASIC. The method separately estimates query-image-to-image and query-text-to-image similarities, performing late fusion to upweight images that satisfy both queries, while downweighting those that exhibit high similarity with only one of the two. Each individual similarity is further improved by a set of components that are simple and intuitive. BASIC sets a new state of the art on *i*-CIR but also on existing CIR datasets that follow a semantic-level class definition. Project page: https://vrg.fel.cvut.cz/icir/.

[*]Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Composed image retrieval (CIR) combines image-to-image retrieval and text-to-image retrieval. CIR uses a composed query, *i.e*. an image and text, to retrieve images whose content matches both the visual and textual parts of the query. Vision and language models (VLMs) [30, 20, 21, 16, 43] provide the foundation for developing CIR methods, either through further training [25, 8, 1, 26] or in a training-free manner [28, 42, 19]. The use of VLMs, owing to their large-scale pre-training, enables CIR to operate in an open-world setting and compare any kind of visual or textual content. This capability paves the way for novel applications and advanced methods to explore and browse large image collections. However, the main limitation of CIR lies in the lack of appropriate data for both evaluating progress and training models. This work addresses these challenges.

Existing CIR datasets [25, 39, 2] often suffer from poor quality due to their construction process, *i.e*., two similar images are selected automatically and their difference is textually described. This approach incorrectly assumes that such a difference always forms a meaningful text query for retrieval, regardless of the image pair. In contrast, given one image, we first specify a textual modification such that both together form a meaningful composed query. We then identify positive and a large set of hard negative images to construct *i*-CIR, a compact yet challenging dataset. The goal is to retrieve images that depict the same object instance as the image query, under the modification described by the text query. Such an instance-level object class definition is missing from existing datasets and is identified as a limitation by prior work [33]. By integrating diverse object types and modification types, *i*-CIR accurately reflects a wide range of real-world use cases.

CIR methods that rely on further training on top of VLMs require large amounts of training triplets in the form of (query image, query text, positive image), which are challenging to obtain at scale. As a result, training is typically performed on small sets of manually labeled triplets [38, 25, 1, 4], or on automatically generated triplets obtained through crawling [44, 17, 24] or synthetic data generation [10]. However, these automated methods significantly compromise triplet quality, and in all cases, the diversity of visual object types and textual modifications remains limited compared to the variety present in VLM pre-training, thereby restricting generalization ability. Instead, we develop a training-free Baseline Approach for SurprIsingly strong Composition, BASIC, which fully leverages existing VLM capabilities and remains adaptable to future advances.

BASIC separately computes the similarity with respect to each query component and performs fusion inspired by the classical Harris corner principle [12]; both responses must be jointly high, rather than just one of the two. The image-to-image dot product similarity is enhanced through a projection learned not in the image space, but in the text representation space. This is facilitated by a large language model that provides common object names and typical textual modifications. The aim is to increase distinctiveness regarding object variations, *i.e*., to better represent image objects beyond other visual cues, while achieving invariance to textually described modifications, so that the same object is retrieved despite variations. Interestingly, this projection can also be equivalently applied solely on the query side, enabling user-specific or application-specific customization. The text-to-image dot product similarity is refined via a query-time contextualization process designed to bridge the distribution gap between the text inputs seen during VLM pre-training and the text queries used at inference. Our contributions are summarized as follows:

- We introduce *i*-CIR, a new evaluation dataset for CIR, meant to retrieve images containing the same particular object as the visual query under modifications defined by the text query.
- We introduce BASIC, a training-free approach leveraging pre-trained VLMs for class-level or instance-level CIR that is based on image-to-image and text-to-image similarities, without the need to update the database embeddings.
- BASIC sets a new state of the art on *i*-CIR and across existing class-level CIR datasets.

2 Related work

Methods. While early methods like TIRG [38], CIRPLANT [25] and CLIP4CIR [1] rely on supervised training with annotated triplets, recent efforts in zero-shot CIR (ZS-CIR) avoid triplet supervision and fall into three main categories. *Textual-inversion* methods (*e.g*., Pic2Word [33], SEARLE [2], ISA [6], LinCIR [11]) map the reference image to a pseudo-text token, which is then

composed with the modification text in the language domain and processed by a vision-language model. *Pseudo-triplet* approaches (*e.g.*, TransAgg [24], HyCIR [17], CompoDiff [10], CoVR-2 [37], MCL [22]) generate synthetic training data using LLMs [23] and image generative models [31], either from caption-editing strategies or from natural web-based image pairs. *Training-free methods* (*e.g.*, WeiCom [28], FreeDom [42], CIReVL [19], GRB [35], WeiMoCIR [40],) leverage off-the-shelf VLMs [30, 20, 21] and LLMs [36] to perform CIR without any additional training by either recasting it as text-based retrieval or fusing visual and textual embeddings directly via weighted sums or geometric interpolations.

Datasets. Key benchmarks include FashionIQ [39] (77k images, 30k triplets across three fashion sub-tasks) and CIRR [25] (22k images, 37k triplets), both criticized for label ambiguity, high false-negative rates, and text-only shortcuts [2, 15], and recently refined by [15]. CIRCO [2] (1k queries over 120k COCO-unlabeled distractors, 4.53 targets/query) extends this paradigm with more diverse negatives. Four additional domain-conversion datasets-ImageNet-R [13] (30k stylized images of 200 classes in four style domains), MiniDomainNet [46] (140k images of 126 classes in four domains), NICO++ [45] (89k images of 60 categories in six contexts), and LTLL [7] (500 images of 25 locations)-explore class-level retrieval under style or context shifts. Concurrent to our work, ConCon-Chi [32] introduces an image–caption benchmark for personalized concept–context understanding, designed for text-to-image generation, retrieval, and editing. Their *concepts* correspond to our *instances*, while their *contexts* parallel the modifications expressed by our *text queries*.

3 *i*-CIR dataset

3.1 Overview and structure

We introduce the *i*-nstance-level Composed Image Retrieval (*i*-CIR) evaluation dataset. Following the instance-level class definition [29, 41], we group all visually indistinguishable objects, *i.e.* the same particular object, into a single class. For example, a class may correspond to (i) a concrete physical entity, such as the Temple of Poseidon, or (ii) a fictional yet visually distinctive character or object, such as Batman. In practice, if a human observer can confidently state that multiple visual manifestations represent the *same object*, they belong to the same instance-level class.

Given a *composed query* (q^v, q^t) consisting of an image query q^v depicting a particular object, also referred to as an object instance or simply *instance*, and a text query q^t describing a modification, the goal is to rank a database of images such that those depicting the same instance under the requested modification appear at the top. We refer to these images as *composed positives* or simply *positives*. For each composed query, we consider the following types of hard negative images: (i) *visual hard negative*: depicts an identical or visually similar object as q^v but does not match the textual modification q^t, (ii) *textual hard negative*: matches the semantics of q^t but depicts a different instance, typically from the same semantic category, (iii) *composed hard negative*: nearly matches both query parts, while one of the two may be identically matched, *i.e.* depicts an object similar/identical to q^v with semantics similar to q^t, or an object similar to q^v with semantics identical/similar to q^t. Examples are shown in Figure 1 and Figure 3.

All types of negatives, including non-hard ones, are treated equally during evaluation. However, we include a significant number of hard negatives in our dataset to create a challenging yet manageable benchmark that supports future research. There are n^v image queries for the same instance combined with n^t text queries that are combined to construct $n^v n^t$ composed queries (values of n^v and n^t vary per instance). Unlike typical retrieval benchmarks that use a single common database for all queries, we employ the same image database for all $n^v n^t$ composed queries of an instance, but a different database for queries of other instances. This design ensures scalable and error-free labeling, avoiding the impracticality of verifying each database image as positive or negative for every query.

3.2 Collection and curation

The dataset construction process combines human input with automated image retrieval[2]. Our aim is to curate, for each instance, composed queries, sets of corresponding positives, and a well-structured set of challenging hard negatives, with all *i*-CIR images sourced from the LAION [34] dataset.

[2]We perform *image-to-image* and *text-to-image* retrieval using dot product search based on image and text representations obtained from OpenAI CLIP [30].

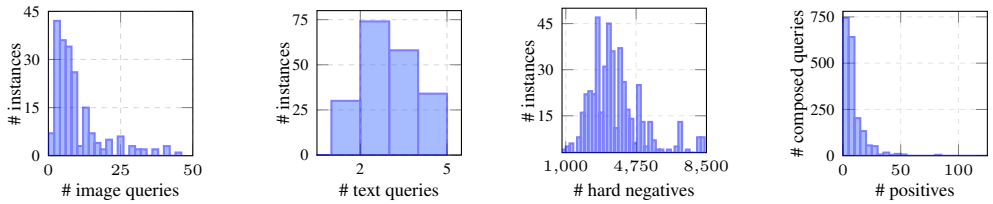


Figure 2: i-*CIR statistics*. From left to right: Number of (a) image queries, (b) text queries, and (c) hard negatives per instance; (d) composed positives per composed query.

The process for each instance begins by defining the object instance, *e.g.* Temple of Poseidon, and selecting semantically meaningful modifications, *e.g.* ‘‘`at sunset`’’, while avoiding implausible ones, *e.g.* ‘‘`with snow`’’. We then create *seed images* and *seed sentences* to serve as queries for retrieving neighbors from LAION, which collectively form a *candidate image pool* that includes potential queries, positives, and (hard) negatives.

Seed images: 2 to 5 high-quality images depicting (i) the object instance, *e.g.* the Temple of Poseidon, or (ii) a composed positive, *e.g.* the Temple of Poseidon at sunset. These images are gathered from web searches in Creative Commons repositories or personal photo collections. The neighbors retrieved from LAION are categorized as visual and composed hard negatives for cases (i) and (ii), respectively.

Seed sentences: Textual descriptions of (i) the instance (*e.g.* ‘‘`Temple of Poseidon`’’), (ii) another object of the same category (*e.g.* ‘‘`Ancient Greek Temple`’’), (iii) rephrased versions of defined modifications (*e.g.* ‘‘`a photo of dusk`’’), (iv) the instance under the modifications (*e.g.* ‘‘`Temple of Poseidon at sunset`’’), (v) an object of the same semantic category under the modification (*e.g.* ‘‘`an Ancient Greek Temple at sunset`’’). The neighbors retrieved from LAION are classified as visual, textual, and composed hard negatives for cases (i & ii), (iii), and (iv & v), respectively.

After building the candidate image pool, automated filtering removes low-resolution, watermarked, or duplicate content using perceptual hashing and resolution checks. Annotators then manually inspect the remaining images to identify composed *positives* per composed query. Unmarked images are considered *negatives*. *Visual hard negatives* are associated with all composed queries of an instance, while *textual* and *composed hard negatives* are associated only with the specific composed query from which (or from whose text query) they were derived. Finally, annotators manually select images within the visual hard negatives to serve as *image queries*. All images that were neither filtered out from the candidate image pool nor selected as queries form the database for this instance. Positives and hard negatives associated with a composed query are negatives for another composed query.

To avoid bias towards/against CLIP-based methods, seed images are discarded and not included in *i*-CIR, while seed sentences do not include the exact phrasing of a text query.

3.3 Statistics and visualisations

Figure 2 summarizes key per-instance and per-query statistics in *i*-CIR. We include 202 object instances and 750 K images in total. Each instance has 1–46 image queries (195 with >1, median: 6) and 1–5 text modifications (median: 2), yielding 1,883 composed queries overall. Queries can be categorized either by their visual part (the object instance) or by their textual part (the modification). Each composed query has 1–127 positives (median: 5) and each instance's database contains 951–10,045 hard negatives (median: 3,420), creating a challenging retrieval benchmark. Figure 3 illustrates a set of randomly chosen pairings from the categorization: for each of eight visual–textual category combinations, we show the image query, composed positive, visual hard negative, textual hard negative, and composed hard negative. These visualisations highlight the rich diversity of *i*-CIR, both in terms of the wide array of visual categories (*e.g.*, landmarks, products, fictional characters, tech gadgets) and the broad spectrum of textual modifications (*e.g.*, viewpoints, attributes, contexts, additions), setting our benchmark apart from existing CIR datasets.

3.4 Shortcomings of existing benchmarks

Commonly used CIR datasets include CIRR [25], FashionIQ [39], CIRCO [2], and ImageNet-R [13]. CIRR, FashionIQ, and CIRCO share a common limitation: their construction relies on an automated process to select two similar images, guided by either textual or visual similarity. These images form the image query and the positive pair. Due to the nature of the image sources, either there is no obvious relation of such selected pairs or the relationship is typically at semantic level only,

Figure 3: *Visualization of visual and textual category examples from* i-*CIR*. For each of the eight randomly chosen category pairings (a–h), we display: the image query, the text query, the composed positive, the visual hard negative, the textual hard negative, and the composed hard negative.

rather than at instance level. Subsequently, a language-based description is generated to capture the difference between the two images. However, the lack of concrete differences between the images, often coupled with their low relevance, results in descriptions that are either poor representations of meaningful text queries or inadequate components of a composed query. In many cases, the text query alone suffices to describe the positive image, making the image query redundant. Moreover, these datasets exhibit a paucity of challenging negatives and a substantial rate of false negatives in their ground-truth annotations [2], inflating reported performance. We present such cases in the supplementary material.

Domain-conversion benchmarks such as ImageNet-R [13], NICO++ [45], and MiniDomainNet (MiniDN) [46] extend CIR to style or context shifts (*e.g.*, ‘‘`photo`’’→‘‘`cartoon`’’) but define positives by semantic class membership rather than object identity, lacking instance-level granularity. LTLL [7] is the sole existing instance-level domain-conversion dataset, but it is extremely limited in scale (500 images of 25 locations, two domains) and provides only binary ‘‘`archive`’’ *vs.* ‘‘`today`’’ modifications. Furthermore, these benchmarks offer very narrow categorical variation—FashionIQ is confined to fashion items, LTLL to a two-way temporal shift, and the domain sets to domain changes only. These semantic-level definitions, small scale, minimal textual variation, and weak negative mining in prior benchmarks motivate the creation of *i*-CIR.

4 A surprisingly strong baseline

In the task of *composed image retrieval* (CIR), we are given an image query $q^v \in \mathcal{X}^v$ and a text query $q^t \in \mathcal{X}^t$, where $\mathcal{X}^v$ is the image input space and $\mathcal{X}^t$ is the text input space. The goal is to retrieve images x^v from a database $X = \{x_1^v, \ldots, x_n^v\} \subset \mathcal{X}^v$ that are visually relevant to the image query and reflect the modifications specified by the text query. Features are extracted using a pre-trained visual encoder $\phi^v : \mathcal{X}^v \to \mathbb{R}^d$ and text encoder $\phi^t : \mathcal{X}^t \to \mathbb{R}^d$, *e.g.*, CLIP, which map image and text queries to a shared embedding space of dimension d. Image-to-image and image-to-text similarities are computed via dot product of the corresponding features.

The proposed training-free method, called BASIC, is based on the assumption that both modalities in the composed retrieval query encode complementary information that jointly contribute to the retrieval objective. This makes the composed retrieval task analogous to performing a logical conjunction over the two modalities: we seek results that are simultaneously relevant to both the image and the text [3].

[3]While this assumption holds well for standard composite tasks (*e.g.*, this image and the concept “winter”), it may not apply in tasks where one modality dominates (*e.g.*, purely textual transformations) or where the text query is highly entangled with image content (*e.g.*, CIRR-like datasets). In such cases, the benefits of the proposed mechanisms may diminish.

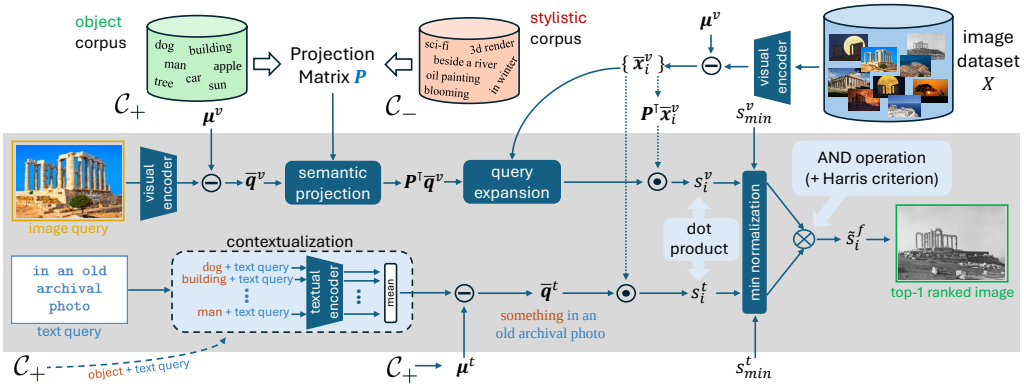


Figure 4: *Overview of our training-free composed image retrieval method* BASIC. Given a query image and text, we apply centering and semantic projection, guided by corpora $C+$ and $C-$, to suppress irrelevant dimensions. The text is contextualized using caption-like prompts. Both modalities are scored against the database with min-based normalization and fused via a multiplicative "AND" operation regularized by a Harris-like criterion to retrieve jointly relevant results.

Following the aforementioned assumption, we compute the similarity per modality and then combine them. We improve the representation per modality by removing modality-specific noise and spurious correlations that can interfere with their effective combination. In practice, visual features may be entangled with background clutter, image composition, or dataset-specific styles, whereas textual features can reflect lexical biases or corpus-level drift. Figure 4 presents a high-level overview of BASIC, which consists of a sequence of conceptually simple yet effective steps that progressively filter out the aforementioned noise and modality-specific artifacts from the image and text features.

Centering for bias removal. We remove modality-specific biases by subtracting mean features. These means typically capture low-level regularities unique to each modality: the average image feature reflects general visual patterns, while the average text feature captures common linguistic patterns. Subtracting them helps isolate semantic content from distributional bias. In particular, after extracting the image features $\mathbf{q}^v = \phi^v(q^v) \in \mathbb{R}^d$ and text features $\mathbf{q}^t = \phi^t(q^t) \in \mathbb{R}^d$, also from the database images, we subtract a precomputed image feature mean $\boldsymbol{\mu}^v \in \mathbb{R}^d$, and a text feature mean $\boldsymbol{\mu}^t \in \mathbb{R}^d$, respectively. The centered features are

$$\bar{\mathbf{q}}^v = \mathbf{q}^v - \boldsymbol{\mu}^v = \phi^v(q^v) - \boldsymbol{\mu}^v \quad \text{and} \quad \bar{\mathbf{q}}^t = \mathbf{q}^t - \boldsymbol{\mu}^t = \phi^t(q^t) - \boldsymbol{\mu}^t. \tag{1}$$

To ensure scalability and generalization, we compute $\boldsymbol{\mu}^v$ using a large external dataset X^v, *e.g.*, LAION [34]. Similarly, we calculate the text mean $\boldsymbol{\mu}^t$ on a predefined textual corpus which contains content-relevant concepts (see next step).

Projection onto semantic subspace. We aim to transform the image features to retain information related to the main objects, while suppressing information related to image styles, object domains, or background setting, *i.e.* that correspond to text query modifications of common use cases. This is achieved by projecting into a lower-dimensional subspace derived from text CLIP features. To construct this projection, we use two textual corpora: C_+, a *object corpus* containing object-oriented terms (*e.g.*, "building", or "dog"), and C_-, a *stylistic corpus* containing terms related to style, viewing conditions or contextual setting (*e.g.*, "cartoon", "aerial view", or "in a cloudy day"). Inspired by [27], we compute a weighted contrastive covariance matrix as follows :

$$\mathbf{C} = (1-\alpha)\mathbf{C}_+ - \alpha\mathbf{C}_-, \quad \text{where} \quad \mathbf{C}_\pm = \frac{1}{|C_\pm|} \sum_{x \in C_\pm} (\phi^t(x) - \boldsymbol{\mu}^t)(\phi^t(x) - \boldsymbol{\mu}^t)^\top \tag{2}$$

and α is an empirically determined hyperparameter. We extract the top-k eigenvectors of $\mathbf{C} \in \mathbb{R}^{d \times d}$ to form a projection matrix $\mathbf{P} \in \mathbb{R}^{d \times k}$. The eigenvectors capture directions with high variance in $\mathbf{C}_+$ and small in $\mathbf{C}_-$, emphasizing object-specific cues while suppressing style-related variation. We then project the centered image features $\bar{\mathbf{x}}^v = \phi(x^v) - \boldsymbol{\mu}^v \in \mathbb{R}^d$, either query or database, as $\mathbf{P}^\top \bar{\mathbf{x}}^v \in \mathbb{R}^k$. Note that the corpora C_+ and C_- need not match the domain of the retrieval database. Even generic corpora for C_+, *e.g.*, class names from ImageNet, yield performance improvements, as the captured directions are semantically rich and broadly transferable.

Image query expansion. In the literature, image retrieval performance, recall in particular, has been shown to be significantly improved by query expansion [3, 9]. The proposed method benefits from

applying the optional step of query expansion using the image query. Following [9], the original feature of the image query is enhanced by a weighted combination of the top-ranked database features that it retrieves. The weights are an increasing function of the corresponding similarities.

Contextualization of text queries. CLIP is trained primarily on natural language captions and full sentences. As a result, using single-word text queries (*e.g.*, ‘‘`sculpture`’’) or sentence parts (*e.g.*, ‘‘`during sunset`’’) constitute out-of-distribution input and may produce text features that are not well-aligned with image features. To address this, we introduce a *contextualization* step that enriches such textual queries with additional terms. Let q^t be a raw text query (*e.g.*, ‘‘`sculpture`’’). We generate multiple caption-like queries by combining q^t with elements from the subject corpus C_+. We add a random set of terms before (*e.g.*, ‘‘`dog during the sunset`’’) and after (*e.g.*, ‘‘`sculpture dog`’’) the text query. These composed phrases are embedded using CLIP's text encoder, centered, and averaged. This operation yields a more robust textual feature that better reflects how CLIP interprets concepts in natural language (*e.g.*, ‘‘`[something] during the sunset`’’).

Score normalization and fusion. The final step is to combine similarities from the two modalities to rank the database items. Given the centered image query embedding $\bar{\mathbf{q}}^v \in \mathbb{R}^d$ and the contextualized centered text query (either original, or expanded) embedding $\bar{\mathbf{q}}^t \in \mathbb{R}^d$, we compute similarities to the centered embedding $\bar{\mathbf{x}}^v \in \mathbb{R}^d$ of a database image $x^v \in X$ as:

$$s^v = \left\langle \mathbf{P}^\top \bar{\mathbf{x}}^v, \mathbf{P}^\top \bar{\mathbf{q}}^v \right\rangle \quad \text{and} \quad s^t = \left\langle \bar{\mathbf{x}}^v, \bar{\mathbf{q}}^t \right\rangle . \tag{3}$$

To reflect the complementary nature of the modalities, we fuse the two scores by multiplication: $s = s^v s^t$. However, due to modality imbalance and differences in representation ranges, one modality can disproportionately dominate the final score.

Min-based normalization. To mitigate range imbalances, an affine re-scaling of the similarities s in each modality is performed. The empirical minimum $s_{\min} < 0$ of the dot product in (3) is used, so that $s_{\min}$ is mapped to 0 and 0 is mapped to 1:

$$\tilde{s} \;=\; (s - s_{\min})/|s_{\min}|.$$

We apply this to both s^v and s^t using predefined statistics for $s^v_{\min}$ and $s^t_{\min}$, estimated on an external dataset. See supplementary material for details.

Fused similarity with Harris criterion. Finally, we fuse the normalized scores using multiplication and a regularizer inspired by the Harris corner detector. The final score is:

$$\tilde{s}^f = \tilde{s}^v \tilde{s}^t - \lambda(\tilde{s}^v + \tilde{s}^t)^2.$$

The first term rewards items that are jointly relevant to both modalities. The second term penalizes items where only one modality is highly activated, thereby suppressing false positives from unbalanced queries. The scalar λ controls the trade-off and is fixed across all experiments.

Computational complexity. The strengths of our approach stem from its simplicity and efficiency. The entire pipeline is (deep-network) training-free and is composed of operations that scale linearly or sub-linearly with the dataset size, since similarity computation over the dataset items is a simple inner product and can be efficiently handled by existing libraries, *e.g.*, FAISS [18]. The proposed similarity computation efficiently operates over a stored database of original CLIP representations. The similarity s^v is efficiently computed as follows

$$s^v = \left\langle \mathbf{P}^\top(\mathbf{x}^v - \boldsymbol{\mu}^v), \mathbf{P}^\top(\mathbf{q}^v - \boldsymbol{\mu}^v) \right\rangle = \left\langle \mathbf{x}^v, \mathbf{P}\mathbf{P}^\top(\mathbf{q}^v - \boldsymbol{\mu}^v) \right\rangle - \left\langle \boldsymbol{\mu}^v, \mathbf{P}\mathbf{P}^\top(\mathbf{q}^v - \boldsymbol{\mu}^v) \right\rangle ,$$

where the first term is a dot-product computed over the unaltered database features and the second term is a query dependent constant. Thus, all computation related to centering and projection can be computed on-the-fly on the query side. This is valuable, since the mean and the projection matrix can be alternated (*e.g.* with specific knowledge of the domain) without touching the stored index. This makes our method particularly well-suited for large-scale deployments, requiring no adaptation, no fine-tuning, and no backpropagation.

5 Experiments

5.1 Experimental setup

Datasets and evaluation protocol. We evaluate BASIC on our proposed *i*-CIR as well as four composed image retrieval benchmarks: ImageNet-R, MiniDN, NICO++, and LTLL. Retrieval

performance is measured using the standard mean Average Precision (mAP) metric. Average Precision (AP) is computed per query by averaging the precision values at the ranks of all relevant items in the retrieval list. The mean Average Precision (mAP) is then obtained by averaging AP over all queries, providing a global measure of retrieval effectiveness that accounts for the order of relevant results. For *i*-CIR, we report the *macro-mAP* over instances, defined by first computing mAP per instance and then taking the mean of these per-instance mAPs across all instances.

Baselines and competitors. We include four simple baselines. "Text" scores each database image $x^v \in X$ by $\langle \phi^t(q^t), \phi^v(x^v) \rangle$; "Image" scores by $\langle \phi^v(q^v), \phi^v(x^v) \rangle$; "Text + Image" combines the similarities by summation; "Text $\times$ Image" by product. We also benchmark BASIC against state-of-the-art zero-shot composed image retrieval methods: WeiCom [28], Pic2Word [33], CompoDiff [10], CIReVL [19], SEARLE [2], MCL [22], MagicLens [44], CoVR-2 [37], and FreeDom [42]. All methods use CLIP with ViT-L/14 [5], whereas CompoDiff employs the larger CLIP ViT-G/14.

BASIC. For fair comparison, we also use CLIP [30] ViT-L/14 [5]. We set $k = 250$ components for PCA, $\lambda = 0.1$ for the Harris criterion and $\alpha = 0.2$. These values were fixed once on a small privately owned development set, named *i*-CIR $_{\text{dev}}$. The corpora C_+ and C_- were automatically generated using ChatGPT [14]. The statistics $s^v_{\min}$ and $s^t_{\min}$ were computed over a synthetically generated dataset constructed using Stable Diffusion [31] with automatically created prompts. More details are included in the supplementary material.

5.2 Experimental results

Per-category performance. In Figure 5 we report the per-category performance of selected baselines and competitors on *i*-CIR split by the a) primary visual and (b) textual categories of the queries.

In Figure 5(a), BASIC ranks first in six of the eight visual categories, delivering particularly large margins on fictional (47.8% *vs*. 31.1% for SEARLE), mobility (45.8% *vs*. 29.3% for MagicLens), and technology (30.6% *vs*. 23.0% for Text × Image). It also leads on product (33.7% *vs*. 26.7% for MagicLens), landmark (39.3% *vs*. 35.0% for MagicLens), and art (38.0% *vs*. 35.0% for MagicLens). The only exceptions are fashion, where MagicLens edges out at 25.6% *vs*. 22.0% for BASIC, and household, where MagicLens peaks at 29.1%; BASIC is second at 22.4%. In contrast, the other methods show uneven strengths.

Figure 5(b) further confirms the consistency of BASIC. BASIC dominates projection (53.1% *vs*. 31.1% for MagicLens), appearance (48.8% *vs*. 36.8% for SEARLE), and domain (39.3% *vs*. 31.1% for MagicLens). It also leads on viewpoint (47.8% *vs*. 40.1% for MagicLens) and attribute (26.3% *vs*. 24.1% for MagicLens). BASIC is second on context (35.6% *vs*. 36.4% for MagicLens) and addition (24.0% *vs*. 28.2% for MagicLens).

Comparison with SOTA. We further evaluate the performance of BASIC against all considered baselines and state-of-the-art CIR methods across five datasets, including *i*-CIR. Results are shown in Table 1. As observed, BASIC consistently outperforms all competing methods across the board. Runtime comparisons are provided in the supplementary material.

Table 1: Average mAP (%) comparison across datasets. †: without query expansion.

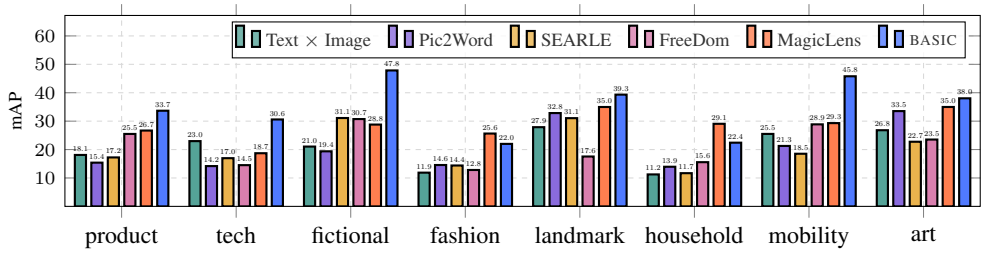
Method	ImageNet-R	NICO++	MiniDN	LTLL	*i*-CIR
Text	0.74	1.09	0.57	5.72	3.01
Image	3.84	6.32	6.66	16.49	3.04
Text + Image	6.21	9.30	9.33	17.86	8.20
Text × Image	7.83	9.79	9.86	23.16	17.48
WeiCom	10.47	10.54	8.52	26.60	18.03
Pic2Word	7.88	9.76	12.00	21.27	19.36
CompoDiff	12.88	10.32	22.95	21.61	9.63
CIReVL	18.11	17.80	26.20	32.60	18.66
SEARLE	14.04	15.13	21.78	25.46	19.90
MCL	8.13	19.09	18.41	16.67	19.89
MagicLens	9.13	19.66	20.06	24.21	27.35
CoVR-2	11.52	24.93	27.76	24.68	28.50
FreeDom	29.91	26.10	37.27	33.24	17.24
FreeDom †	25.81	23.24	32.14	30.82	15.76
BASIC	**32.13**	**31.65**	**39.58**	**41.38**	31.64
BASIC †	27.54	28.90	35.75	38.22	**34.35**

Note. For *i*-CIR we report *macro-mAP*

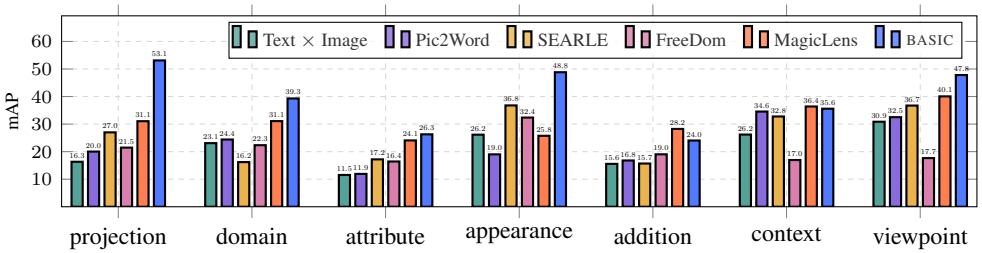
5.3 Ablation studies

BASIC components. Table 2 presents a detailed ablation study on the contribution of each component of BASIC across all evaluated datasets. Starting with a simple Text×Image baseline, we progressively add the components of BASIC, highlighting the cumulative benefits of each module.

Centering provides a notable boost across most datasets (17.48% $\rightarrow$ 28.33% on *i*-CIR) with the exception of LTLL, likely due to its narrow focus on landmarks. *Normalization and Harris fusion* further enhance retrieval, as demonstrated by their removal, with min normalization being especially



(a) *Visual categories.* mAP (%) averaged over instances in the same primary-level visual category.



(b) *Textual categories.* mAP (%) averaged over instances in the same primary-level textual category.

Figure 5: *Performance comparison on* i*-CIR per primary category of queries.* (a) Visual, (b) Textual.

Table 2: Ablation study reporting average mAP (%) across datasets. Each row progressively adds or removes components of the proposed method: mean centering (*Centering*), min-based normalization (*Min Norm.*), Harris criterion (*Harris*), text contextualization (*Context.*), semantic projection (*Proj.*), and query expansion (*Q. Exp.*). The first row (no component applied) corresponds to Text × Image.

Centering	Min Norm.	Harris	Context.	Proj.	Q. Exp.	ImageNet-R	NICO++	MiniDN	LTLL	*i*-CIR
✗	✗	✗	✗	✗	✗	7.66	9.26	9.48	19.78	17.48
✓	✗	✗	✗	✗	✗	12.16	9.95	12.16	16.93	28.33
✓	✓	✗	✗	✗	✗	12.06	17.20	17.72	22.20	27.30
✓	✓	✓	✗	✗	✗	16.21	15.06	17.79	29.70	28.42
✓	✓	✓	✓	✗	✗	18.61	15.34	21.01	33.74	33.48
✓	✓	✓	✓	✓	✗	27.54	28.90	35.75	38.22	**34.35**
✓	✓	✓	✓	✗	✓	17.31	13.96	21.22	22.42	31.78
✓	✓	✓	✗	✓	✓	26.18	30.61	33.64	34.50	25.85
✓	✓	✗	✓	✓	✓	30.75	29.82	38.85	40.65	31.61
✓	✗	✗	✓	✓	✓	24.50	22.74	29.65	19.36	30.75
✓	✓	✓	✓	✓	✓	**32.13**	**31.65**	**39.58**	**41.38**	31.64

critical, since its absence causes a significant drop. Harris consistently contributes moderate gains. *Text contextualization* is also important. Its removal results in a substantial performance decline, particularly on datasets requiring nuanced language understanding (31.64% → 25.85% on *i*-CIR). On the image side, *semantic projection* accounts for the majority of the performance gain in many cases, serving as a key enhancement. *Query expansion* offers additional improvements, particularly on category-level datasets, though it leads to performance decrease in *i*-CIR. Note that some components depend on the presence of others to be effective (*e.g.*, projection assumes centered features, Harris step requires min-normalized scores).

Table 3: mAP(%) on each dataset using different negative corpora. The first column lists the evaluation datasets.

	Negative Corpora Source					
Eval. Dataset	none	generic	Imagenet-R	NICO++	MiniDN	LTLL
Imagenet-R	30.15	32.13	**33.22**	30.74	31.17	30.91
NICO++	30.67	**31.65**	30.84	31.17	30.67	31.20
MiniDN	38.64	39.38	39.34	38.75	**39.58**	38.68
LTLL	41.80	41.24	41.33	43.39	42.09	**43.98**
i-CIR	31.51	**31.64**	31.61	31.20	31.32	31.06

Controlling semantic projection. Table 3 shows the effect of omitting C_{-}, using a generic negative corpus, or using a dataset-specific corpus (generated via ChatGPT) designed to reflect the domain variability of ImageNet-R, NICO++, MiniDN, and LTLL. Results indicate that leveraging application-related knowledge can improve performance, particularly compared to omitting C_{-}. This idea is further discussed in the supplementary material.

***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. Using the performance of Text × Image baseline as a reference (17.48%), we find that more than 40M distractor images are required for this baseline to

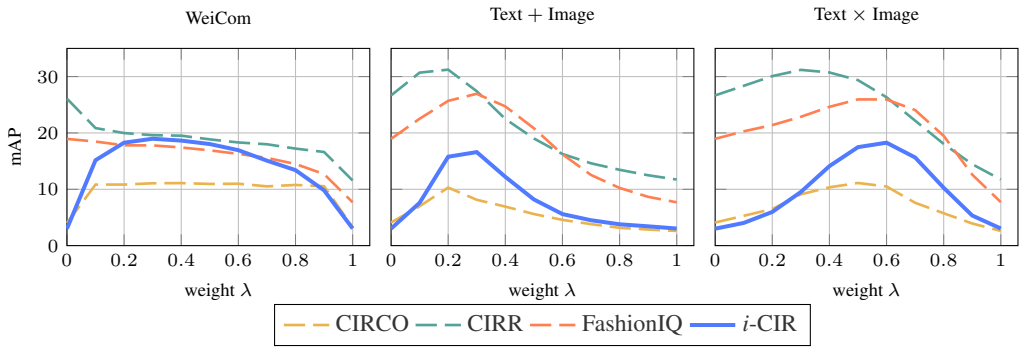


Figure 6: *Modality bias via weight sweeps.* mAP(%) *vs*. fusion weight λ for three simple methods, where λ=0 is text-only and λ=1 is image-only. Compositional datasets should peak at interior λ and exceed both endpoints. *i*-CIR shows strong interior optima and large gains over the best uni-modal baseline; CIRR and FashionIQ peak at λ=0, indicating text dominance.

reach a similarly low performance. Note that the performance measured using unlabeled LAION images as negatives is only a lower bound, due to the inevitable presence of false negatives. This is four orders of magnitude larger than the 3.7K database images per query, on average, that *i*-CIR uses, or 1.5 orders of magnitude larger than the 750K database images used among all queries; the latter defines the experimental processing cost. More analysis is provided in the supplementary material.

***i*-CIR: Truly compositional.** A dataset that requires *composition* should reward combining text and image, not either modality alone. To diagnose this, we sweep a mixing weight $\lambda \in [0, 1]$ between text-only (λ=0) and image-only (λ=1) similarity for three simple fusion methods (WeiCom, Text + Image, Text $\times$ Image), and plot mAP as a function of λ (Figure 6). For each method we compute the *composition gain* Δ: the difference between the peak value of the curve and the best uni-modal endpoint, and then average Δ across the three methods. On *i*-CIR, the average composition gain is large: $+14.9$ mAP ($+490\%$ relative to the best uni-modal baseline), with peaks occurring at interior λ—clear evidence that both modalities must work together. By contrast, it shrinks to $+3.0$ mAP ($+11\%$) on CIRR, $+5.0$ mAP ($+26\%$) on FashionIQ, and $+6.8$ mAP ($+167\%$) on CIRCO. Moreover, the highest uni-modal performance of CIRR and FashionIQ is always when ($\lambda = 0$), *i.e.*, text-only. Thus legacy datasets reward composition only marginally, whereas *i*-CIR demands genuine cross-modal synergy; BASIC is designed for the latter scenario.

6 Conclusions

We introduced *i*-CIR, an instance-level benchmark for composed image retrieval with *explicit hard negatives* (visual, textual, and composed). It fills a long-standing gap by providing an ambiguity-free evaluation suite that rewards composition rather than single-modality shortcuts. We also presented BASIC, a simple, efficient, *training-free* method that compares favorably to both training-based and training-free baselines across benchmarks. BASIC is built from a few transparent components, whose combination delivers strong accuracy, transfers well, and exhibits broad hyperparameter plateaus. We hope *i*-CIR becomes a reliable target for assessing genuinely compositional retrieval, and that the simplicity of BASIC catalyzes adoption and further advances.

Acknowledgments. This work was supported by the Czech National Recovery Plan—CEDMO 2.0 NPO (MPO 60273/24/21300/21000) provided by the Ministry of Industry and Trade of the Czech Republic; the EU Horizon Europe programme MSCA PF RAVIOLI (No. 101205297) and HERON - Hellenic Robotics Center of Excellence (No. 101136568); the National Recovery and Resilience Plan (NNRP) "Greece 2.0"/NextGenerationEU project "Applied Research for Autonomous Robotic Systems" (MIS5200632); and Czech Technical University in Prague (SGS23/173/OHK3/3T/13 and institutional Future Fund). We acknowledge VSB – Technical University of Ostrava, IT4Innovations National Supercomputing Center, Czech Republic, for awarding this project (OPEN-33-67) access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium, through the Ministry of Education, Youth and Sports of the Czech Republic via the e-INFRA CZ project (ID: 90254). We also acknowledge the OP VVV project "Research Center for Informatics" (CZ.02.1.01/0.0/0.0/16 019/0000765).

References

[1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, 2022.

[2] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, 2023.

[3] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007.

[4] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. 2022.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[6] Yongchao Du, Min Wang, Wengang Zhou, Shuping Hui, and Houqiang Li. Image2sentence based asymmetrical zero-shot composed image retrieval. *arXiv preprint arXiv:2403.01431*, 2024.

[7] Basura Fernando, Tatiana Tommasi, and Tinne Tuytelaars. Location recognition over large time lags. *Computer Vision and Image Understanding*, 139:21–28, 2015.

[8] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *CVPR*, 2022.

[9] Albert Gordo, Filip Radenovic, and Tamara Berg. Attention-based query expansion learning. In *ECCV*, 2020.

[10] 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.

[11] 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.

[12] Chris G. Harris and Mike Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, Manchester, UK, 1988.

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021.

[14] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[15] 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.

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[17] Yingying Jiang, Hanchao Jia, Xiaobing Wang, and Peng Hao. Hycir: Boosting zero-shot composed image retrieval with synthetic labels. *arXiv preprint arXiv:2407.05795*, 2024.

[18] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. In *arXiv*, 2017.

[19] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *arXiv*, 2023.

[20] 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.

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[22] Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan Kankanhalli. Improving context understanding in multimodal large language models via multimodal composition learning. In *ICML*, 2024.

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[24] Yikun Liu, Jiangchao Yao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Zero-shot composed text-image retrieval. *arXiv preprint arXiv:2306.07272*, 2023.

[25] 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.

[26] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5753–5762, 2024.

[27] James Oldfield, Christos Tzelepis, Yannis Panagakis, Mihalis Nicolaou, and Ioannis Patras. Parts of speech–grounded subspaces in vision-language models. *Advances in Neural Information Processing Systems*, 36:2700–2724, 2023.

[28] Bill Psomas, Ioannis Kakogeorgiou, Nikos Efthymiadis, Giorgos Tolias, Ondrej Chum, Yannis Avrithis, and Konstantinos Karantzalos. Composed image retrieval for remote sensing. In *IGARSS*, 2024.

[29] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018.

[30] 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.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[32] Andrea Rosasco, Stefano Berti, Giulia Pasquale, Damiano Malafronte, Shogo Sato, Hiroyuki Segawa, Tetsugo Inada, and Lorenzo Natale. Concon-chi: Concept-context chimera benchmark for personalized vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22239–22248, 2024.

[33] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *CVPR*, 2023.

[34] 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.

[35] Shitong Sun, Fanghua Ye, and Shaogang Gong. Training-free zero-shot composed image retrieval with local concept reranking. *arXiv preprint arXiv:2312.08924*, 2023.

[36] 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.

[37] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Varol Gül. Covr-2: Automatic data construction for composed video retrieval. *TPAMI*, 2024.

[38] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*, 2019.

[39] 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.

[40] Ren-Di Wu, Yu-Yen Lin, and Huei-Fang Yang. Training-free zero-shot composed image retrieval via weighted modality fusion and similarity. In *International Conference on Technologies and Applications of Artificial Intelligence*, pages 77–90. Springer, 2024.

[41] Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne van Noord, and Giorgos Tolias. The met dataset: Instance-level recognition for artworks. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021.

[42] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. *arXiv preprint arXiv:2303.09833*, 2023.

[43] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

[44] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*, 2024.

[45] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *CVPR*, 2023.

[46] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *ICLR*, 2021.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize our key contributions: the creation of an instance-level composed image retrieval dataset (*i*-CIR) with challenging hard negatives, the proposal of a training-free CIR method (BASIC) that leverages existing VLMs, and the demonstration of state-of-the-art performance on *i*-CIR and multiple benchmarks. All claims are supported by detailed methodology and experiments in Sections 3–5, as also in supplementray material.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly acknowledge in Sec. 3 (footnote) that our logical-AND assumption may break down when one modality dominates (e.g., pure style edits or highly entangled image–text pairs like CIRCO). We point readers to additional method and dataset limitations in the supplementary material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper contains no formal theorems or theoretical results requiring proof; it is entirely empirical and algorithmic.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We give exhaustive details of our dataset construction (including LAION search and retrieve, filtering and annotation protocols) and our method (all pre-processing steps, corpora generation, hyper-parameters k, λ, α, normalization statistics), and we commit to releasing both the dataset and code with scripts to reproduce every result.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Both *i*-CIR dataset and code will be made publicly available through our project page https://vrg.fel.cvut.cz/icir/.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental section of both the main paper and supplementary material is exhaustive and detailed. It specifies all the details needed to understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our method is training-free, so it does not involve sources of randomness such as weight initialization, optimization, or data shuffling.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Technical and implementation details are included in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We use only publicly available, license-annotated sources (*e.g.*, LAION-derived URLs and rights-cleared repositories), and we respect licenses and site Terms of Service. All images in *i*-CIR were reviewed by trained annotators; inappropriate, copyrighted, or privacy-sensitive content was removed. In categories where people are inherently present (*e.g.*, apparel), we automatically pixelate faces and perform spot checks; no raw PII is released. *i*-CIR is an evaluation-only benchmark, distributed under CC-BY-NC-SA with an explicit prohibition on surveillance/biometric or other privacy-invasive uses. We publish a misuse policy, provide a "Report misuse/PII" channel, honor takedown requests, and reserve the right to revoke access for violations. Our method is training-free and does not scrape private data, minimizing environmental and privacy risks. No human-subjects research was conducted (IRB not applicable). Overall, collection, curation, release, and documentation follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper explicitly discusses benefits and risks. On the positive side, instance-level composed retrieval can support cultural-heritage search (GLAM), assistive-vision use cases, product provenance, and reproducible evaluation of compositional models. On the negative side, we analyze dual-use pathways—including surveillance/profiling, indirect "object-of-interest" tracking (*e.g.*, via distinctive belongings), fine-tuning our techniques on face/plate corpora, and misuse of crawl scripts—and we describe harms from both correct and incorrect system behavior. However, we exhaustively outline concrete mitigations too.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We release *i*-CIR as an *evaluation-only* benchmark under CC-BY-NC-SA with an explicit ban on surveillance/biometric and other privacy-invasive uses; access is governed by a misuse policy with revocation. Images are curated with a privacy-first process: annotators preferentially exclude PII, and in categories where people are intrinsic (*e.g.*, apparel) we retain images but *exhaustively pixelate* visible faces before release; we also filter watermarks, near-duplicates, and low-quality items, followed by manual spot checks. Any search/crawl scripts are released under the same restrictive license with hard-coded keyword blocks and documentation discouraging sensitive-content collection (IRB review recommended for modifications). The project page provides a prominent "Report misuse / PII" channel; we commit to prompt review, content takedown, and access revocation when warranted. We will periodically red-team object-level re-identification risks and update the release if failure modes are found. No identification models or person-level embeddings are released; our method is training-free and not tailored to biometric tasks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We build on publicly available assets and cite them in the paper and project page. For data, we rely on LAION *metadata* (CC-BY 4.0) and keep each image under

its original source license. For models/code, we use public VLMs and toolchains under their original licenses and cite them: CLIP (MIT), OpenCLIP (Apache-2.0), and, for anonymization, InsightFace (MIT). Our own release is evaluation-only under CC-BY-NC-SA and does not alter upstream terms; LICENSE files and attributions are included in our repo.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release *i*-CIR with a comprehensive datasheet detailing collection protocols, licensing, annotation guidelines, dataset statistics, and limitations; our code repository includes installation instructions, example scripts, configuration files for all experiments, and explicit CC-BY-NC-SA license information.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: We did not use crowdsourcing; all annotations were performed by salaried institutional staff. The supplemental material includes annotation guidelines/instructions. Annotators received domain and ethics training, and we ran weekly QA spot-checks and inter-annotator-agreement audits. Compensation exceeds the legal minimum (> 80% of a first-year PhD stipend with full social-security coverage). No participants were recruited, no demographic attributes were collected, and PII was removed/redacted; thus this work does not constitute human-subjects research, but we nevertheless followed our institution's ethics policies throughout.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not constitute human-subjects research: no participants were recruited or interacted with; annotators were salaried staff performing routine labeling; no demographic or behavioral data were collected; and all images were sourced from public datasets with PII removed/pixelated prior to release. Under these conditions, IRB (or equivalent) review is not required per common definitions and institutional policies.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.