

UNION: A Lightweight Target Representation for Efficient Zero-Shot Image-Guided Retrieval with Optional Textual Queries

Hoang-Bao Le¹, Allie Tran¹, Binh T. Nguyen², Liting Zhou¹, Cathal Gurrin¹

¹ADAPT Centre, School of Computing, Dublin City University

²Ho Chi Minh University of Science, Vietnam National University

Dublin, Ireland

bao.le2@mail.dcu.ie, {liting.zhou, cathal.gurrin}@dcu.ie

Abstract—Image-Guided Retrieval with Optional Text (IGROT) is a general retrieval setting where a query consists of an anchor image, with or without accompanying text, aiming to retrieve semantically relevant target images. This formulation unifies two major tasks: Composed Image Retrieval (CIR) and Sketch-Based Image Retrieval (SBIR). In this work, we address IGROT under low-data supervision by introducing UNION, a lightweight and generalizable target representation that fuses the image embedding with a null-text prompt. Unlike traditional approaches that rely on fixed target features, UNION enhances semantic alignment with multimodal queries while requiring no architectural modifications to pretrained vision-language models. With only 5,000 training samples—from LlavaSCo for CIR and Training-Sketchy for SBIR—our method achieves competitive results across benchmarks, including CIRCO mAP@50 of 38.5 and Sketchy mAP@200 of 82.7, surpassing many heavily supervised baselines. This demonstrates the robustness and efficiency of UNION in bridging vision and language across diverse query types. Our code is available at [URL].

Index Terms—Composed Image Retrieval, Sketch-Based Image Retrieval, Zero Shot, Cross-Modal Retrieval, Image Guided Retrieval with Optional Text

I. INTRODUCTION

Composed Image Retrieval (CIR) and Sketch-Based Image Retrieval (SBIR) are two key tasks in content-based image retrieval that involve retrieving a target image based on an input query. In CIR, the query consists of a reference image and a textual modification describing how the target should differ [1]–[3], while SBIR uses a sketch to retrieve a matching natural image [4], [5]. Although these tasks appear different, they share a common structure: an image-driven query with optional auxiliary language input. We unify them under the general formulation of **Image-Guided Retrieval with Optional Text (IGROT)**, where the query always contains an anchor image and may be optionally augmented with descriptive text. IGROT reflects a practical retrieval setting that accommodates both visual-only and vision-language queries within a shared framework.

Despite substantial progress in CIR, developing large-scale, high-quality datasets remains challenging. Creating CIR benchmarks requires carefully curated triplets of reference image, modification text, and target image, which is labor-

intensive and often domain-specific. As a result, most datasets suffer from limited coverage, ambiguous language, and lack of diversity in compositional relations. To reduce this reliance on manual annotation, some recent works adopt zero-shot or weakly supervised paradigms [6]–[9], often using millions of synthetic or web-mined triplets. However, this introduces its own trade-offs: massive data generation is computationally expensive and may contain noisy or unnatural descriptions.

To address these limitations, we construct **LlavaSCo**, a compact and high-quality dataset derived from LaSCo [10], consisting of approximately 5,000 triplets enhanced with detailed captions generated by the vision-language model LLaVA [11]. This small-scale dataset enables effective CIR training under limited supervision while maintaining semantic richness. Complementarily, for the SBIR setting, we create the **Training-Sketchy** dataset from Sketchy [12], which consists of 5,000 sketch-natural image pairs covering diverse object categories, allowing us to validate our method across both language-augmented and visual-only retrieval scenarios.

While prior CIR research has primarily focused on improving the fusion of the query image and modification text to produce a rich compositional embedding, comparatively little attention has been given to how the target image is represented—which in most works is simply the visual feature extracted directly from a pretrained vision-language model, such as CLIP or BLIP, without further adaptation. Most recent methods [9], [10], [13], [14] compare the query against fixed target features directly extracted from pretrained vision-language encoders such as CLIP [15] or BLIP [16]. However, these fixed embeddings may not be well-aligned with the compositional semantics of the query, especially in fine-grained or relational changes.

To address this, we introduce **UNION**, a simple yet effective target feature representation. Instead of comparing queries to frozen image embeddings, UNION computes a composed target feature by combining the visual representation of the target image with a null-text prompt, passed through a lightweight Transformer and MLP. This design allows the target representation to reside in the same vision-language embedding space as the query, enhancing semantic alignment and improving

contrastive training dynamics. Importantly, UNION does not require modifications to the backbone encoder and supports image-only and image-text retrieval.

In this paper, our contributions are summarised as follows:

- We demonstrate that strong CIR performance can be achieved using only 5,000 high-quality triplets from LaSCo, refined into our LlavaSCo dataset with detailed captions generated by LLaVA, significantly reducing the dependence on large-scale annotations.
- We show that our method generalises to the SBIR setting by treating sketches as image-only queries and applying the same architecture, achieving strong performance with minimal adaptation.
- We propose **UNION**, a novel and efficient target feature representation that combines visual and null-text inputs, improving semantic compatibility with fused queries without requiring changes to the pretrained vision-language backbone.

II. RELATED WORK

A. Zero-Shot Image-Guided Retrieval with Optional Text (IGROT)

The task of Image-Guided Retrieval with Optional Text (IGROT) unifies Composed Image Retrieval (CIR) and Sketch-Based Image Retrieval (SBIR) by allowing queries that contain an anchor image with or without accompanying text. While supervised CIR and SBIR methods often require large-scale annotated triplets, recent efforts aim to reduce supervision via zero-shot or weakly supervised approaches.

a) Composed Image Retrieval (CIR): To mitigate the cost of manual annotations in CIR, several works have proposed training strategies that eliminate the need for annotated triplets. Pic2Word [17] introduced the Zero-Shot CIR (ZS-CIR) setting by learning from weakly labeled image-caption pairs and unlabeled image collections. Other methods focus on scaling data through synthetic triplet generation. For example, MagicLens [7] created 36.7M triplets from web images and PaLM2-generated captions [18], while CC-CoIR [19] constructed 3.3M triplets from Conceptual Captions [20]. CoLLM [9] leveraged large language models (LLMs) like Claude 3 Sonnet¹ to generate high-quality relational captions for 3.4M image pairs. In contrast, LaSCo [10] proposed a lightweight method using GPT-3 [21] and data roaming to collect 360k image-text triplets. Recently, LoGra-Med [22] reduced the dependence on large-scale datasets by constructing a smaller but more informative training set, using only 10% of the original data while preserving performance. Following this, our work departs from these scale-intensive strategies by using only 5,000 triplets from LaSCo. We enhance this subset with detailed image captions generated by LLaVA [11], resulting in a compact yet semantically rich dataset we name **LlavaSCo**.

b) Sketch-Based Image Retrieval (SBIR): SBIR has also seen interest in zero-shot generalisation. Earlier methods such as SAKE [23] focused on preserving discriminative semantic features across modalities. Hybrid fusion [24] and adapter-based structural alignment [25] further reduced domain gaps. DCDL [26] used causal disentanglement to improve cross-domain representation, and ZSE-SBIR [27] aligned sketches and photos via local patch correspondences for explainable matching. However, these methods rely exclusively on visual cues, which can limit the model’s ability to distinguish subtle semantic differences or relational concepts—particularly when the query includes nuanced modifications that are better captured through language. In contrast, we introduce **language as an auxiliary modality** in SBIR by pairing sketch queries with a fixed textual prompt during training. This allows us to unify CIR and SBIR under the IGROT framework, using the same model and training strategy to support both query types.

B. Target Representation in Vision-Language Retrieval

While much of the zero-shot CIR literature focuses on improving query fusion ([13], [14]), or expanding training data ([9], [28], [10]), the target representation has received relatively little attention. Most approaches compare the composed query against fixed target embeddings obtained directly from vision-language models such as CLIP [15] or BLIP [16]. These static features may not align well with the compositional semantics of the query, especially in fine-grained or cross-modal scenarios. To address this, we propose the **UNION** feature, which fuses the target image embedding with a null-text prompt using a lightweight Transformer-MLP stack. This produces a semantically enriched representation that resides in the same embedding space as the composed query, improving retrieval quality without modifying the backbone model or requiring additional supervision.

III. METHODOLOGY

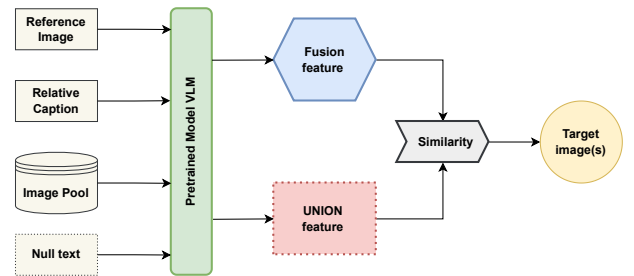


Fig. 1: An Overview of Model Architecture in Composed Image Retrieval. Normally, the models do not have Null Text and UNION Feature and directly rank the similarity score of Fusion feature with Image Pool’s embedded feature.

In this section, first, we present TransAgg [13] as a representative example, to provide an overview of the ZS-CIR paradigm. Following this, we present a detailed explanation of the proposed UNION feature combining the pool’s image

¹<https://www.anthropic.com/claude/sonnet>

and null text. For the ZS-CIR task, we introduce **LlavaSCo**, an updated version of LaSCo [10], which improves both training quality and performance on standard benchmarks. For the ZS-SBIR task, we construct **Training-Sketchy**, a compact training set derived from Sketchy [12], containing 5,000 high-quality samples.

A. Overview

In Composed Image Retrieval (CIR), the goal is to retrieve one or more target images T that satisfy a user query composed of a reference image R and a textual description C specifying the desired transformation. A common modeling approach formulates this task as learning a fusion function $\mathcal{F}_{rc} \in \mathbb{R}^{B \times D}$ with $r \in R$ and $c \in C$, which maps the image-text pair into a joint embedding space. The resulting query representation is then compared against the embedded target image feature $\mathbf{e}_{t_i} \in \mathbb{R}^{B \times D}$, typically using cosine similarity.

To train this representation, we adopt the Batch-Based Classification (BBC) loss [6], which encourages each fused query \mathcal{F}_{rc} to be most similar to its corresponding target \mathbf{e}_t and dissimilar to other targets in the batch. Formally, for a batch of size B , the loss is defined as:

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^B \log \left[\frac{\exp(\text{sim}(\mathcal{F}_{r_i c_i}, \mathbf{e}_{t_i})/\tau)}{\sum_{j=1}^B \exp(\text{sim}(\mathcal{F}_{r_i c_i}, \mathbf{e}_{t_j})/\tau)} \right], \quad (1)$$

where $\text{sim}(a, b) = \frac{a^\top b}{\|a\| \|b\|}$ denotes cosine similarity, and τ is a temperature scaling parameter.

B. UNION: Unified Target Representation

In standard retrieval setups, models typically compare a multimodal query embedding with a fixed image-only target embedding \mathbf{e}_{t_i} . However, in Composed Image Retrieval (CIR), the query consists of a joint representation of image and text, leading to a modality mismatch when compared to a unimodal target feature. This asymmetry makes it difficult for the model to learn fine-grained semantic alignments, especially when the modification involves subtle relational or attribute-level changes. Moreover, the lack of textual conditioning in target features hinders the model’s ability to reason about abstract or compositional semantics.

To bridge this gap, we propose the **UNION** feature—a unified and semantically enriched target representation that incorporates both the visual and latent textual context. Specifically, we concatenate the target image embedding \mathbf{e}_{t_i} with a null-text embedding \mathbf{e}_η , where η is an empty string passed through a pretrained vision-language model (e.g., CLIP or BLIP). The null-text embedding introduces a neutral linguistic prior that implicitly brings the target closer to the multimodal space used for queries, without injecting external textual content.

As shown in Figure 2, the concatenated feature

$$\mathbf{e}_{t\eta} = [\mathbf{e}_t; \mathbf{e}_\eta] \in \mathbb{R}^{B \times 2 \times D},$$

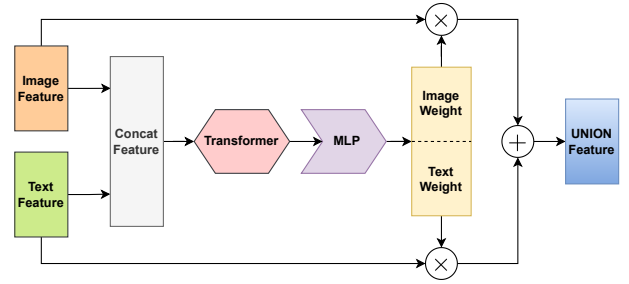


Fig. 2: UNION architecture combining all images to be retrieved and Null Text.

where B is the batch size and D is the feature dimension, is passed through a lightweight Transformer encoder (following the architecture of T5 [29], similar to UNIIR [30]). The Transformer processes the sequence and outputs a tensor of the same shape $\mathbb{R}^{B \times 2 \times D}$, which is then pooled (e.g., using average or the first token) to produce a single feature vector per instance $f^{t\eta} \in \mathbb{R}^{B \times D}$. This vector is passed through a Multi-Layer Perceptron (MLP) to compute the adaptive weight:

$$f^{t\eta} = \text{Transformer}(\mathbf{e}_{t\eta}) \in \mathbb{R}^{B \times 2 \times D} \quad (2)$$

$$w_t = \text{MLP}(\text{Pool}(f^{t\eta})) \in \mathbb{R}^{B \times D}, \quad w_\eta = 1 - w_t \quad (3)$$

The final UNION feature is a learned interpolation between the image and null-text features:

$$\mathcal{U}_i = w_t \cdot \mathbf{e}_t + w_\eta \cdot \mathbf{e}_\eta$$

During inference, UNION continues to use the same null-text conditioning, which does not require any additional user input or textual description, ensuring consistent behavior across CIR and SBIR, including sketch-only queries. This structure improves retrieval quality by aligning the target embeddings more closely with the fused query representations and supports plug-and-play integration with various pre-trained backbones.

To train the model, we simply replace the embedding of the raw target image \mathbf{e}_{t_i} in the loss function (Equation 1) with the UNION feature \mathcal{U}_i , resulting in the updated objective:

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^B \log \left[\frac{\exp(\text{sim}(\mathcal{F}_{r_i c_i}, \mathcal{U}_i)/\tau)}{\sum_{j=1}^B \exp(\text{sim}(\mathcal{F}_{r_i c_i}, \mathcal{U}_j)/\tau)} \right] \quad (4)$$

C. Dataset Construction

1) *LlavaSCo: Caption Refinement for Stronger Alignment.*: While LaSCo [10] provides a large-scale dataset for CIR, we observe that its relative captions often lack specificity or semantic clarity, weakening the learning signal for tasks requiring precise compositional understanding. This is particularly problematic for our UNION-based target representation, which benefits from high-quality text-to-image alignment.

To mitigate this, we enhance a subset of LaSCo using LLaVA [11], a vision-language model capable of generating

detailed captions. For each triplet, we generate a refined target image caption and append it to the original relational caption to produce a more informative instruction. We refer to this enhanced subset (5,000 training triplets) as **LlavaSCo**, which serves as the training data for our CIR experiments. An example is shown in Figure 3. With the generated caption, the instruction that shifts the image input into the retrieved image is less ambiguous and well-described.

Additional details and examples of the caption generation process are included in the Appendix V.

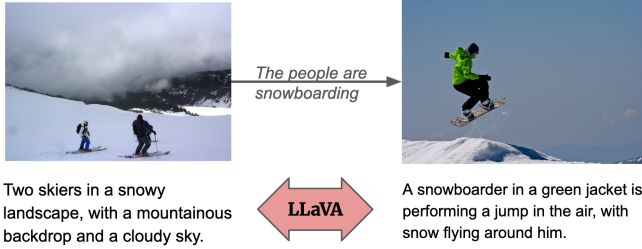


Fig. 3: LLaVA Caption for the target image and also the reference image.

2) *Training-Sketchy: Constructing SBIR Triplets*: As a case study within the IGROT framework, we focus on Sketch-Based Image Retrieval (SBIR), where the query is a sketch and the goal is to retrieve a corresponding real-world image. Although prior ZS-SBIR work often constructs separate training sets for each benchmark (e.g., Sketchy, TU-Berlin, and QuickDraw), our approach uses a unified training set. Specifically, we use the Sketchy dataset [12], which provides sketch-image pairs across 125 categories, and we build a small, curated dataset—Training-Sketchy—from the Sketchy dataset only, and apply it across all Zs-ZS-SBIR benchmarks. We randomly select 50 categories from the official training split and sample 100 sketch-photo pairs per category, resulting in a total of 5,000 training triplets. To align with our vision-language architecture, we associate each sketch query with a fixed instruction:

“a real image of sketch”

This simple prompt allows sketches to be treated similarly to CIR queries that combine an image with text, enabling cross-task training under the IGROT setting. We refer to this dataset as **Training-Sketchy**, which we use to train our model on SBIR alongside CIR data from LlavaSCo.

IV. EXPERIMENT AND DISCUSSION

A. Evaluation Datasets and Metrics

a) *Zero-Shot Composed Image Retrieval*: In the training stage, we use the LlavaSCo dataset—our enhanced version of LaSCo [10]—which shares the same index set. We benchmark our method on three widely-used CIR datasets: FashionIQ [1], which includes over 2,005 triplets across three fashion categories (Dress, Shirt, and Tootie) and a retrieval pool of 5,179 images; CIR [2], comprising 4,148 image-caption queries and 2,316 target images; and CIRCO [3], a large-scale

benchmark with 800 queries and 123,403 candidate images. Following prior work [7], we evaluate retrieval performance using Recall@K for FashionIQ and CIR, and mean Average Precision (mAP) for CIRCO. Crucially, during both training and inference, our approach leverages a null-text prompt when constructing target features using the UNION representation. Based on prior work [7], we adopt Recall at K ($R@K$) for FashionIQ and CIR, and mean Average Precision (mAP) for CIRCO as the evaluation metric.

b) *Zero-Shot Sketch-Based Image Retrieval*: We train the model with the fixed instruction for each sketch query and utilise null text $c = “”$ for inference stage. For testing, we evaluate the model on three benchmarks: Sketchy [4] has 12,694 queries dividing into 21 classes from ImageNet-1k and a 12,694-image pool; TUBerlin [5] consists of 30 categories with 2,400 sketches and a corresponding index set has 27,989 images; QuickDraw [12] includes 92,291 queries covering 30 classes and a 54,146-sized index set. Following prior work [26], we report mAP and Precision at K (Prec@K) for each dataset.

To demonstrate the efficiency of the UNION feature, we also consider the other metrics in Appendix V.

B. Implementation Details

Our framework is implemented with Pytorch. We follow the TransAgg setups [13] and use the transformer-based 2 layer fusion module with 8 heads and GELU activation. For visual and text encoders, we utilise three popular pretrained models as OpenAI CLIP-B/32 and CLIP-L/14 [15], and BLIP with ViT-B [16]. In the UNION architecture, we use a 2-layer transformer architecture similar to the T5 Transformer [29] with 2 layers, but 8 attention heads for CLIP_{base} and BLIP, and 12 attention heads for CLIP_{large} with each head having 64 dimensions. In the loss function, we set $\tau = 0.01$ similar to TransAgg [13] settings. For a fair comparison, all experiments use 224×224 images to ensure, and additionally, we only choose the results used the aforementioned pre-trained models. To compare with our UNION feature \mathcal{U} , we consider the original target image’s feature \mathbf{e}_t and the sum feature $\mathbf{e}_{t\eta}$ of target image t and null text η .

The model is optimised with AdamW [35] optimiser with a weight decay of $1e^{-2}$. All experiments are conducted with 2 epochs using learning rate $1e^{-4}$ and a batch size of 32 on one NVIDIA A100 80GB. For the generated captions in LlavaSCo, we adapt the vision language model LLaVA-v1.6-Mistral-7B.

C. Main Results

a) *Zero-Shot Composed Image Retrieval*: We compare our model in Table I against various ZS-CIR methods: i-SEARLE [31], CIReVL [32], Pic2Word [17], MLLM-I2W [33], LinCIR [14], CoLLM [9], MagicLens [7], and TransAgg [13].

Despite being trained on only 5,000 samples from the enhanced LlavaSCo dataset, TransAgg + UNION delivers remarkably competitive performance across all three ZS-CIR benchmarks. On CIRCO, it achieves the best scores in both

Method	Backbone	# Params	# Triplets	FashionIQ (R)		CIRR (R)		CIRCO (mAP)	
				@10	@50	@10	@50	@10	@50
Pic2Word [17]	CLIP-L	429M	3M	24.7	43.7	65.3	87.8	9.5	11.3
i-SEARLE [31]	CLIP-L	442M	205K	29.2	49.5	66.7	88.8	13.6	16.3
CIReVL [32]	CLIP-L	12.5B	-	28.6	48.6	64.9	86.3	19.1	20.9
MLLM-I2W [33]	CLIP-L	-	3M	30.3	50.1	68.4	92.4	-	-
PLI [34]	CLIP-L	428M	695K	35.4	57.4	69.3	89.8	14.2	16.4
LinCIR [14]	CLIP-L	-	5.5M	26.4	46.6	66.9	88.8	13.9	16.2
MagicLens [7]	CLIP-L	465M	36.5M	30.7	52.5	74.4	92.6	30.8	34.4
CoLLM [9]	BLIP-B	-	3.4M	34.6	56.0	78.6	94.2	20.4	23.1
TransAgg [13]	BLIP-B	235M	32K	34.4	55.1	77.9	93.4	32.2	36.2
TransAgg + UNION	BLIP-B	235M	5K	31.9	51.5	77.6	92.9	34.5	38.5

TABLE I: Comparison of our method against baseline on three benchmarks of ZS-CIR task. While we reproduce the results of TransAgg on CIRCO, the others are from the original papers. Red and Blue numbers indicate the best and second-best results.

mAP@10 (34.5) and mAP@50 (38.5), surpassing all other methods, including those trained on datasets with millions of triplets. This result highlights UNION’s effectiveness in improving target representation by aligning better with the fused multimodal query. While it slightly trails in FashionIQ and CIRR benchmarks compared to models trained on much larger datasets (e.g., CoLLM, MagicLens), the performance gap is relatively small (e.g., 31.9 vs. 34.6 on FashionIQ@10), especially considering the drastic reduction in training data size. These findings demonstrate that UNION not only improves semantic alignment and contrastive learning efficiency but also enables high performance under minimal supervision, reinforcing its value for scalable and data-efficient ZS-CIR systems.

b) Zero-Shot Sketch-Based Image Retrieval: We compare our approach with several ZS-SBIR methods using the same backbones: DCDL [26], CAT [36], IVT [37], ZSE-SBIR [27], and MagicLens [7].

Table II compares the performance of our UNION-enhanced TransAgg framework against state-of-the-art ZS-SBIR methods across Sketchy [4], TU-Berlin [5], and QuickDraw [38]. With only 5,000 training pairs, TransAgg + UNION achieves the best results on Sketchy (82.7 mAP@200 and 79.9 Prec@200), improving upon the original TransAgg by +3.1 mAP and +4.1 Prec@200. This improvement is particularly notable given that prior methods like DCDL and CAT rely on up to 57K training pairs yet still underperform.

On QuickDraw, UNION reaches the second-best mAP (33.4), close to DCDL (33.6) while providing a substantial +8.0 Prec@200 gain over TransAgg (41.5 vs. 33.5). Similarly, on TU-Berlin, UNION achieves 51.0 mAP, outperforming most baselines and showing a +6.0 mAP improvement over TransAgg, though slightly below DCDL, which benefits from larger-scale training (57K–236K pairs) and task-specific optimisation.

Overall, these results confirm that UNION features effectively enrich target representations and narrow the modality gap between sketches and real images. The consistent improvements in both precision and mAP demonstrate that UNION not only enhances overall ranking quality but also improves top-k retrieval accuracy. This makes it a robust and resource-friendly solution for ZS-SBIR, especially under

limited supervision scenarios.

D. Qualitative Results for ZS-IGROT

Figure 4 presents qualitative retrieval examples from the CIRCO validation set under the ZS-CIR setting, where evaluation is performed using Recall@ K , with K corresponding to the number of ground-truth target images per query. We observe that the model enhanced with the UNION feature retrieves target images that more accurately reflect the intended textual modifications compared to the baseline. In cases involving subtle or attribute-level changes—such as alterations in object color, quantity, or context—UNION facilitates more precise semantic alignment. In contrast, the original fixed feature often yields visually similar but semantically mismatched results. These qualitative results further support our claim that UNION enhances compositional reasoning and improves fine-grained retrieval, even in limited supervision scenarios.

E. Ablation Study

a) Effect of Target Representations and Captions in ZS-CIR: Figure 5 presents a heatmap comparing average ZS-CIR performance (the average of all the metrics used in Table I) across backbones (CLIP-B, CLIP-L, BLIP) and target feature types (original, sum, UNION), with and without enhanced captions from LlavaSCo. Firstly, incorporating LlavaSCo captions consistently boosts retrieval performance for all backbones—highlighting the importance of high-quality, detailed textual supervision in learning fine-grained visual transformations. For example, BLIP’s average score increases from 35.4 to 48.6 (original), and from 32.1 to 49.8 (UNION).

Secondly, UNION shows clear advantages when combined with strong vision-language models and semantically enriched training data. BLIP with LlavaSCo achieves the best overall score (49.8), indicating that UNION benefits from both expressive backbones and rich caption context. However, when trained without captions, UNION may underperform slightly, e.g., on CLIP-B, it trails both original and sum features, suggesting its reliance on latent multimodal grounding even when explicit text is absent.

The results validate the design of UNION as a semantically adaptive representation, especially effective under caption-rich scenarios, while also demonstrating the synergy between improved data quality and retrieval-aware target representations.

Method	Backbone	# Pairs	Sketchy		TU-Berlin		QuickDraw	
			mAP@200	Prec@200	mAP	Prec@100	mAP	Prec@200
DCDL [26]	CLIP-B	57K/15K/236K	72.6	76.9	63.4	74.1	33.6	29.6
CAT [36]	CLIP-B	57K/15K/236K	71.3	72.5	63.1	72.2	20.2	38.8
IVT [37]	ViT-B	57K/15K/236K	61.5	69.4	55.7	62.9	32.4	16.2
ZSE-SBIR [27]	ViT-L	57K/15K/236K	52.5	62.4	54.2	65.7	14.5	21.6
MagicLens [7]	CLIP-L	36.7M	68.2	75.8	62.9	73.1	15.1	20.4
TransAgg	CLIP-L	5K	79.6	75.8	45.4	68.2	30.1	43.5
TransAgg + UNION	CLIP-L	5K	82.7	79.9	51.0	69.8	33.4	41.5

TABLE II: Comparison of our method against existing frameworks on three benchmarks of ZS-SBIR task. Except MagicLens and Ours, the others are trained on their own training sets. Red and Blue numbers indicate the best and second-best results.

b) Evaluating UNION Feature Effectiveness in Zero-Shot Sketch-Based Image Retrieval: Figure 6 presents an updated evaluation of three target feature types (original, sum and UNION) across the CLIP-B, CLIP-L, and BLIP backbones in the ZS-SBIR task, measured by average mAP. Across all backbones, UNION demonstrates superior or comparable performance, showcasing its effectiveness in learning robust retrieval representations with minimal data.

The most notable improvement is seen with CLIP-L, where UNION achieves the highest average mAP score of 55.7, significantly outperforming both the original (51.7) and the sum (47.3) variants. This indicates that UNION is especially beneficial when paired with a strong and expressive vision-language model, helping the network better align sketch queries with photo targets. The improvement margin here suggests that UNION’s feature fusion better captures semantic relationships between modalities, which is a key challenge in ZS-SBIR.

In the CLIP-B setting, UNION also leads with 38.5, outperforming original (36.9) and sum (33.1), highlighting its ability to compensate for limited representational power in smaller models. Interestingly, in the BLIP setting, while all three variants yield closely matched results, UNION still edges out the others slightly (46.9 vs. 48.7 original and 46.8 sum), showing that its benefit persists even in high-capacity backbones but with diminishing marginal returns.

These findings confirm that UNION enhances target representation quality and retrieval accuracy in ZS-SBIR, particularly for mid-sized models like CLIP-L, where the balance between semantic flexibility and model capacity is most advantageous.

c) Comparative Analysis of Target Feature Types:

Figure 7 compares the performance of three target feature types—original, sum, and UNION—across six datasets spanning both ZS-CIR (FashionIQ, CIRR, CIRCO) and ZS-SBIR (Sketchy, TU-Berlin, QuickDraw). UNION consistently delivers superior or competitive results, highlighting its robustness across modalities and tasks. In ZS-SBIR benchmarks such as Sketchy and TU-Berlin, UNION clearly leads, achieving the highest scores (82.7 and 51.0, respectively), which indicates its strength in handling sketch-image modality gaps through adaptive semantic alignment.

In the CIR domain, UNION matches or slightly surpasses the original in CIRR (78.0 vs. 77.3) and CIRCO (34.5 vs. 33.3), while staying competitive on FashionIQ. Notably, the

sum-based feature consistently underperforms, especially on QuickDraw (20.2 vs. 33.4 with UNION), suggesting that simple additive fusion fails to capture meaningful compositional nuances. These results reinforce that UNION is not only effective in enriching the target representation with latent context, but also generalises well across structured and unstructured retrieval tasks without requiring architectural changes.

d) Limitations: While our approach demonstrates strong performance with limited supervision, it also presents some practical limitations. Firstly, the construction of the LlavaSCo dataset requires running LLaVA-1.6 Mistral to generate captions for 360k reference-target image pairs, which is computationally expensive and time-consuming at scale. Secondly, the inference stage incurs additional latency due to feature processing: generating target embeddings with CLIP-B, CLIP-L, and BLIP across original, sum, and UNION settings takes on average 271 seconds, 334 seconds, and 656 seconds respectively. This highlights a trade-off between representation quality and computational efficiency, which could be further addressed in future work through faster captioning or feature caching strategies.

V. CONCLUSION

In this work, we explore the challenge of **Image-Guided Retrieval with Optional Text** under limited supervision. We introduce **LlavaSCo**, a caption-enhanced dataset constructed from LaSCo [10] using LLaVA-generated descriptions, and show that training on just 5,000 samples from **Training-Sketchy** is sufficient to achieve strong retrieval performance on ZS-SBIR task. Central to our approach is the proposed **UNION** feature, which replaces traditional fixed target embeddings by combining image features with a null-text prompt. This representation improves semantic alignment with the query and enhances contrastive discrimination, all without modifying pretrained vision-language backbones. However, UNION relies on the presence—or generation—of informative captions to fully leverage linguistic cues, and it introduces additional inference overhead when embedding large image pools. Future directions include exploring more efficient captioning strategies and optimising inference time.

APPENDIX A: LLAVASCO CAPTION GENERATION.

We observe that LaSCo triplets often contain weak or generic relational captions that do not clearly convey the transformation from reference to target image. To strengthen this

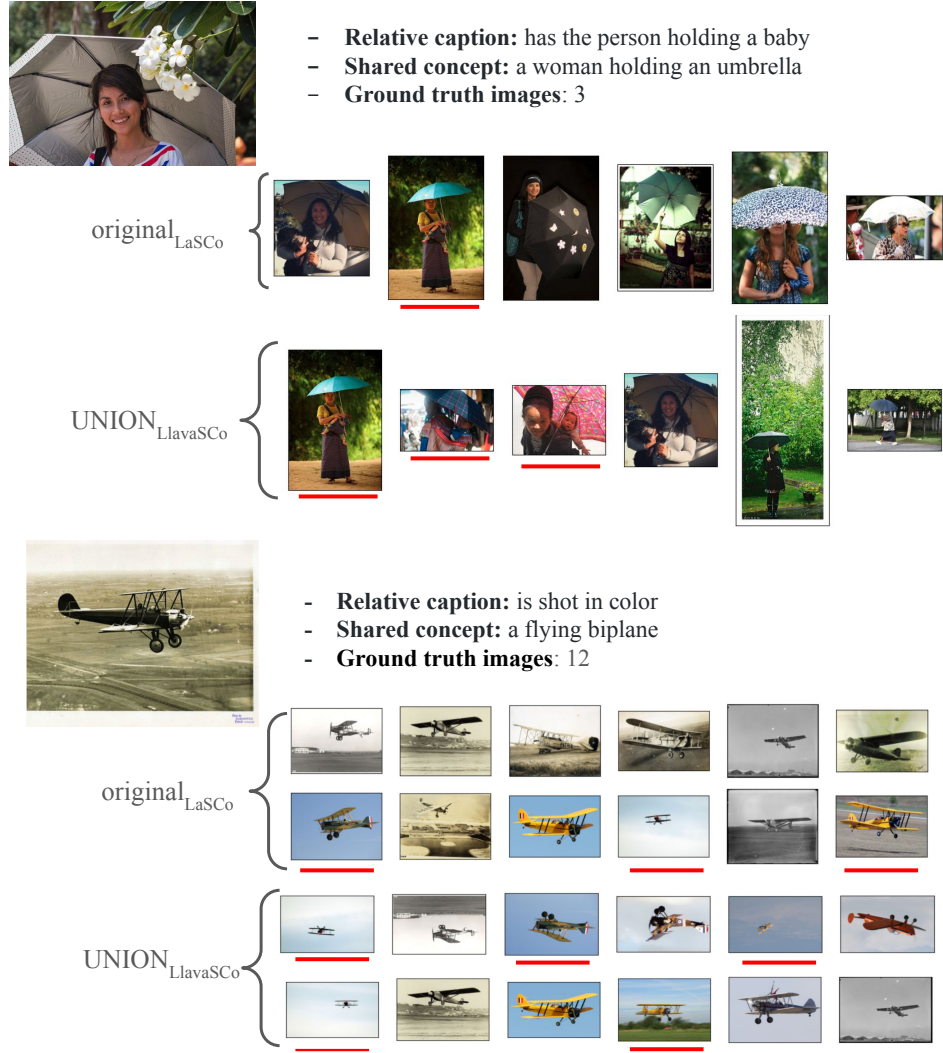


Fig. 4: Qualitative Results on CIRCO validation set . The ground truth images are red-underlined.

link, we enhance each triplet using LLaVA [11], a multimodal model trained to generate fine-grained image descriptions.

For each target image, we generate a caption using the following prompt:

Describe the image in one sentence with details.

We extract only the first sentence to ensure brevity and consistency. The new relative caption is formed as:

$$\text{RelCap}_{\text{new}} = \text{RelCap}_{\text{old}} \text{ with } \text{GenCap}_{\text{target}}.$$

From the full LaSCo dataset (360k triplets), we select a subset of 5,000 refined triplets to form the training set for our CIR experiments. The original validation and test sets are kept unchanged. This refinement helps bridge the modality gap for our UNION representation by improving the textual grounding of each training instance.

APPENDIX B: ADDITIONAL EVALUATION METRICS.

Besides the standard retrieval metrics, we introduce two complementary evaluation criteria to better assess the effectiveness of UNION: **Median Rank (MdR)** and **Mean Average Precision per Ground Truth Number (mAP/GTN)**.

First, Median Rank (MdR) measures the median position of the ground truth image in the ranked retrieval list across all queries; a lower MdR indicates better retrieval precision. Second, we observe that as the retrieval cutoff k increases, Precision scores typically decline. To mitigate this bias, we adopt mAP/GTN, which computes mean average precision where the number of retrieved candidates is set exactly to the number of ground truth images for each query—providing a fairer evaluation of methods in multi-ground-truth settings.

We evaluate these two metrics on the CIRCO validation split and the Sketchy dataset, both of which provide multiple correct target images per query.

As shown in Table III, UNION consistently outperforms

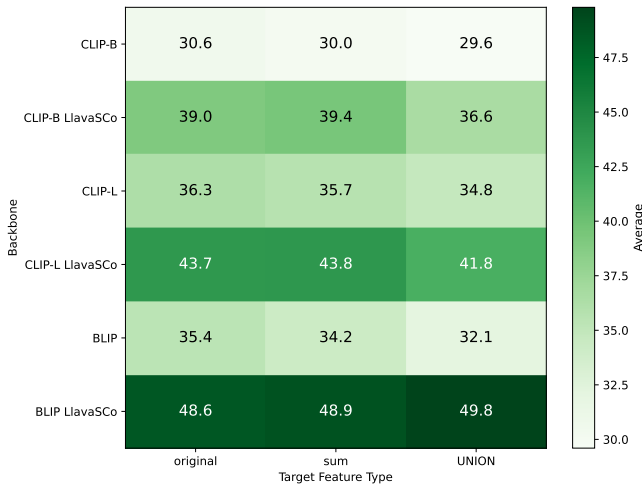


Fig. 5: Heatmap showing average performance across different target feature types (original, sum, UNION) and backbones (CLIP-B, CLIP-L, BLIP), with and without enhanced captions from LlavaSCo, on the ZS-CIR task. UNION consistently improves performance in caption-rich scenarios, especially with stronger backbones like BLIP and CLIP-L, while the benefit diminishes slightly when trained without text.

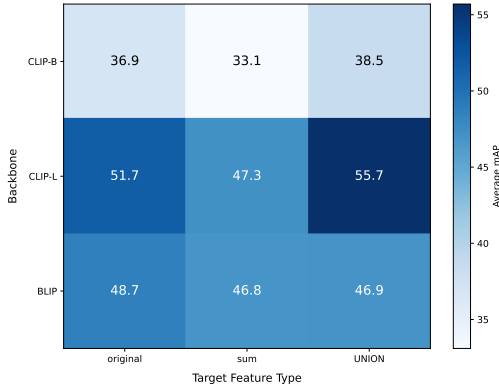


Fig. 6: Heatmap of average mAP performance on the ZS-SBIR task for three target feature types across CLIP-B, CLIP-L, and BLIP backbones. The UNION feature outperforms the sum variant and often surpasses the original, particularly with CLIP-L (48.6) and BLIP (44.4), demonstrating its effectiveness in bridging modality gaps between sketches and real images.

baseline target features on both CIRCO and Sketchy datasets under the new metrics. On CIRCO_{val}, UNION achieves the lowest MdR (5.3) and a highly competitive mAP/GTN (32.2), confirming its ability to retrieve relevant targets with fewer distractors. On Sketchy, UNION yields a substantial improvement, boosting mAP/GTN to 68.8, compared to 59.9 (original) and 57.1 (sum), while also lowering the MdR to 6.1. These results further demonstrate that UNION enhances both ranking quality and retrieval robustness, especially when multiple

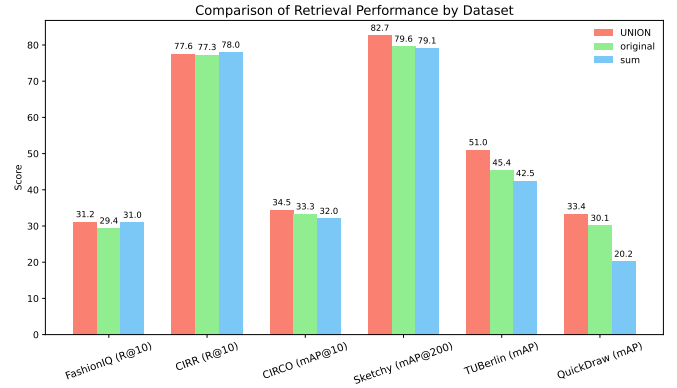


Fig. 7: Comparison between three types of target feature: UNION \mathcal{U} , original \mathbf{e}_t , sum $\mathbf{e}_t + \mathbf{e}_\eta$. We train on LlavaSCo for ZS-CIR task and Training-Sketchy for ZS-SBIR task.

Dataset	Method	MdR ↓	mAP/GTN ↑
CIRCO _{val}	original _{LaSCo}	6.9	24.9
	original _{LlavaSCo}	5.8	32.7
	UNION _{LlavaSCo}	5.3	32.2
Sketchy	original	6.7	59.9
	sum	6.5	57.1
	UNION	6.1	68.8

TABLE III: Performance comparison under Median Rank (MdR) and Mean Average Precision per Ground Truth Number (mAP/GTN) on CIRCO_{val} and Sketchy datasets. UNION consistently improves retrieval ranking and robustness across multiple ground-truth settings.

correct answers exist.

ACKNOWLEDGMENT

This publication has emanated from research supported in part by research grants from Science Foundation Ireland (SFI) under grant numbers SFI/13/RC/2106_P2 and 18/CRT/6223, and co-funded by the European Regional Development Fund.

REFERENCES

- [1] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, “The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback,” *CVPR*, 2021.
- [2] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould, “Image retrieval on real-life images with pre-trained vision-and-language models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 2125–2134.
- [3] A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo, “Zero-shot composed image retrieval with textual inversion,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 338–15 347.
- [4] S. K. Yelamarthi, S. K. Reddy, A. Mishra, and A. Mittal, “A zero-shot framework for sketch based image retrieval,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 300–317.
- [5] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao, “Sketchnet: Sketch classification with web images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1105–1113.
- [6] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, “Composing text and image for image retrieval - an empirical odyssey,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

- [7] K. Zhang, Y. Luan, H. Hu, K. Lee, S. Qiao, W. Chen, Y. Su, and M.-W. Chang, "Magiclens: Self-supervised image retrieval with open-ended instructions," in *The Forty-first International Conference on Machine Learning (ICML)*, 2024, p. to appear.
- [8] G. Gu, S. Chun, W. Kim, H. Jun, Y. Kang, and S. Yun, "Compodiff: Versatile composed image retrieval with latent diffusion," 2024. [Online]. Available: <https://arxiv.org/abs/2303.11916>
- [9] C. Huynh, J. Yang, A. Tawari, M. Shah, S. Tran, R. Hamid, T. Chilimbi, and A. Shrivastava, "Collm: A large language model for composed image retrieval," *arXiv preprint arXiv:2503.19910*, 2025.
- [10] M. Levy, R. Ben-Ari, N. Darshan, and D. Lischinski, "Data roaming and quality assessment for composed image retrieval," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, pp. 2991–2999, Mar. 2024. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/28081>
- [11] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.
- [12] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2862–2871.
- [13] Y. Liu, J. Yao, Y. Zhang, Y. Wang, and W. Xie, "Zero-shot composed text-image retrieval," *arXiv preprint arXiv:2306.07272*, 2023.
- [14] G. Gu, S. Chun, W. Kim, Y. Kang, and S. Yun, "Language-only training of zero-shot composed image retrieval," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [16] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.
- [17] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, "Pic2word: Mapping pictures to words for zero-shot composed image retrieval," *CVPR*, 2023.
- [18] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.
- [19] L. Ventura, A. Yang, C. Schmid, and G. Varol, "Covr-2: Automatic data construction for composed video retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [20] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2556–2565. [Online]. Available: <https://aclanthology.org/P18-1238>
- [21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [22] D. M. H. Nguyen, N. T. Diep, T. Q. Nguyen, H.-B. Le, T. Nguyen, T. Nguyen, N. Ho, P. Xie, R. Wattenhofer, J. Zhou, D. Sonntag, and M. Niepert, "Logra-med: Long context multi-graph alignment for medical vision-language model," 2024. [Online]. Available: <https://arxiv.org/abs/2410.02615>
- [23] Q. Liu, L. Xie, H. Wang, and A. L. Yuille, "Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3662–3671.
- [24] W. Wu, J. Li, Z. Wu, and J. Xu, "Zero-shot sketch-based image retrieval with hybrid information fusion and sample relationship modeling," in *International Conference on Multimedia Modeling*. Springer, 2025, pp. 337–350.
- [25] J. Zhang and J. Tang, "Adapter with textual knowledge graph for zero-shot sketch-based image retrieval," *IEEE Access*, 2025.
- [26] Q. Li, S. Wang, W. Zhang, S. Bai, W. Nie, and A. Liu, "Dcdl: Dual causal disentangled learning for zero-shot sketch-based image retrieval," *IEEE Transactions on Multimedia*, 2025.
- [27] F. Lin, M. Li, D. Li, T. Hospedales, Y.-Z. Song, and Y. Qi, "Zero-shot everything sketch-based image retrieval, and in explainable style," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 23 349–23 358.
- [28] J. Byun, S. Jeong, W. Kim, S. Chun, and T. Moon, "Reducing task discrepancy of text encoders for zero-shot composed image retrieval," 2024. [Online]. Available: <https://arxiv.org/abs/2406.09188>
- [29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [30] C. Wei, Y. Chen, H. Chen, H. Hu, G. Zhang, J. Fu, A. Ritter, and W. Chen, "Uniir: Training and benchmarking universal multimodal information retrievers," in *European Conference on Computer Vision*. Springer, 2024, pp. 387–404.
- [31] L. Agnolucci, A. Baldrati, M. Bertini, and A. Del Bimbo, "isearle: Improving textual inversion for zero-shot composed image retrieval," *arXiv preprint arXiv:2405.02951*, 2024.
- [32] S. Karthik, K. Roth, M. Mancini, and Z. Akata, "Vision-by-language for training-free compositional image retrieval," 2024. [Online]. Available: <https://arxiv.org/abs/2310.09291>
- [33] T. Bao, C. Liu, D. Xu, Z. Zheng, and T. Xu, "MLLM-I2W: Harnessing multimodal large language model for zero-shot composed image retrieval," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Association for Computational Linguistics, 2025.
- [34] J. Chen and H. Lai, "Pretrain like your inference: Masked tuning improves zero-shot composed image retrieval," 2023. [Online]. Available: <https://arxiv.org/abs/2311.07622>
- [35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [36] A. Sain, A. K. Bhunia, P. N. Chowdhury, S. Koley, T. Xiang, and Y.-Z. Song, "Clip for all things zero-shot sketch-based image retrieval, fine-grained or not," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2765–2775.
- [37] H. Zhang, D. Cheng, Q. Kou, M. Asad, and H. Jiang, "Indicative vision transformer for end-to-end zero-shot sketch-based image retrieval," *Advanced Engineering Informatics*, vol. 60, p. 102398, 2024.
- [38] S. Dey, P. Riba, A. Dutta, J. Lladós, and Y.-Z. Song, "Doodle to search: Practical zero-shot sketch-based image retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2179–2188.