Purified Zero-Shot Sketch-Based Image Retrieval

Yang Zhou, Jingru Yang, Jin Wang, Kaixiang Huang, Guodong Lu, and Shengfeng He, *Senior Member, IEEE*

***Abstract*—Sketches, as a new solution in multimedia systems that can replace natural language, are characterized by sparse visual cues such as simple strokes that differ significantly from natural images containing complex elements such as background, foreground, and texture. This misalignment poses substantial challenges for zero-shot sketch-based image retrieval (ZS-SBIR). Prior approaches match sketches to full images and tend to overlook redundant elements in natural images, leading to model distraction and semantic ambiguity. To address this issue, we introduce a distraction-agnostic framework, purified cross-domain matching (PuXIM), which operates on a straightforward principle: masking and matching. We devise a visual-cross-linguistic (VxL) sampler that generates linguistic masks based on semantic labels to obscure semantically irrelevant image features. Our novel contribution is the concept of purified masked matching (PMM), which comprises two processes: (1) *reconstruction*, which compels the image encoder to reconstruct the masked image feature, and (2) *interaction*, which involves a transformer decoder that processes both sketch and masked image features to investigate cross-domain relationships for effective matching. Evaluated on the TU-Berlin, Sketchy, and QuickDraw datasets, PuXIM sets new benchmarks in terms of performance. Importantly, the distraction-agnostic nature of the matching process renders PuXIM more conducive to training, enabling efficient adaptation to zero-shot scenarios with reduced data requirements and low data quality.**

***Index Terms*—Sketch-based image retrieval, zero-shot learning, cross-domain matching.**

I. Introduction

With the rapid development of multimedia services and the explosive growth of image data, efficient image representation, and AI-generated content (AIGC) have become key components of future multimedia communication systems. Previous human-computer interaction modes have relied primarily on natural language [3], [4]. With the widespread use of touchscreen devices, sketches have emerged as a low-barrier form of communication, transcending language barriers and offering greater convenience and controllability. They provide a novel modality for both image retrieval [5], [6], [7] and generation [8], [9], [10]. Sketch-based image retrieval (SBIR) addresses the critical challenge of cross-domain semantic alignment between sketches and images, with applications in e-commerce, information forensics, intelligent transportation systems, and artificial intelligence-generated content (AIGC). SBIR enables the use of sketches as search queries in image databases, offering intuitive and flexible interactions. In e-commerce, SBIR allows users to input sketches for personalized product recommendations. In information forensics, it facilitates the search for potential clues based on witness sketches. For intelligent transportation, sketches can describe vehicle features for identification and localization. In AIGC, sketches provide a more intuitive and controllable interface than does natural language. A more challenging extension of SBIR is zero-shot sketch-based image retrieval (ZS-SBIR), which bridges the domain gap between limited sketch data and natural images to retrieve matches from unseen categories [11], [12], [13].

A critical issue in ZS-SBIR is the substantial discrepancy between the sparse visual cues of sketches and natural images. As shown in Fig. 1(a), the mean number of semantics present in an image for some categories in the popular ZS-SBIR datasets Sketchy [14] and TU-Berlin [15] is greater than 3, and for some categories, it is even greater than 9. Whereas sketches are composed mainly of simple strokes, a sketch commonly contains only a single semantic target and does not have the complex and rich texture information in natural images, as shown in Fig. 1(b).

Despite advancements in ZS-SBIR, current methods still predominantly employ separate encoders for sketches and images to embed them into a shared semantic space, optimizing with losses such as triplet or contrastive loss to learn joint embeddings [16], [17], [18]. These approaches can be roughly divided into two categories, as shown in Fig. 1(c): direct matching and direct matching with interaction. However, matching sketch–image pairs directly without considering semantic redundancy can lead to matching ambiguity. The model is not aware that the object to be matched by the sketch is the hat, the cow, or the shoe (Fig. 1(c)), which distracts the model during optimization and overly focuses on redundant information, thus weakening the matching relationship between real sketch–image pairs and making effective correspondence learning potentially difficult.

To address this, we propose a distraction-agnostic framework, namely, purified cross-domain matching (PuXIM), for

Received 9 October 2024; revised 29 December 2024 and 18 February 2025; accepted 3 May 2025. Date of publication 14 November 2025; date of current version 8 January 2026. This work was supported in part by the National Key R&D Program of China under Grant 2022YFB3303102 and in part by the Robotics Institute of Zhejiang University under Grant K11808 and Grant K11811. The associate editor coordinating the review of this article and approving it for publication was Prof. Jungong Han. *(Corresponding author: Jin Wang.)*

Yang Zhou, Jin Wang, Kaixiang Huang, and Guodong Lu are with the State Key Laboratory of Fluid Power & Mechatronic Systems, Zhejiang University, Hangzhou 310027, China, and also with the Robotics Research Center of Yuyao City, Ningbo 315400, China (e-mail: 22260043@zju.edu.cn; dwjcom@zju.edu.cn; kaixianghuang@zju.edu.cn; lugd@zju.edu.cn).

Jingru Yang is with the School of Computer Science, Carnegie Mellon University, Pennsylvania, PA 15213 USA (e-mail: jingruyang1617@gmail.com).

Shengfeng He is with the School of Computing and Information Systems, Singapore Management University, Singapore 188065 (e-mail: shengfenghe7@gmail.com).

Digital Object Identifier 10.1109/TMM.2025.3632682

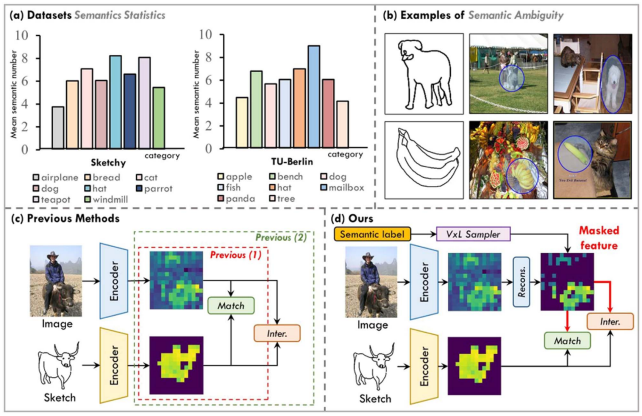

Fig. 1. (a) The semantic statistics on Sketchy and TU-Berlin are derived via a standard DeepLabv3 [1] semantic segmentation model pre-trained on the COCO-Stuff dataset [2] (background instance is not included in the count). (b) Illustrations of the sketch–image pairs with semantic ambiguity. (c) Prior solutions: in *previous method (1)*, sketches are directly matched with complete images. *The previous method (2)* employs direct cross-domain matching for semantic correspondence, both of which encounter background distractions. (d) Our framework mitigates ambiguous semantics via mask reconstruction (recons.) and interaction (inter.), masking 70% of the image features. In previous studies, sketch features often align more closely with ambiguous natural image elements such as "hat" and "shoes". Our approach aims to narrow the gap between the sketch and ground truth ("cow") and increase the distance between sketches and nonground truth semantics for a purified semantic space.

ZS-SBIR. We leverage the semantic labels to directly focus solely on pertinent image regions and eliminate semantic ambiguity. Our visual-cross-linguistic (VxL) sampler, by harnessing vision-and-language pretraining (VLP) technologies [19], [20], generates linguistic masks that identify and omit redundant image pixels, preserving only semantically relevant features.

The method of using our obtained linguistic masks is equally important. Previous approaches, such as hard masked matching [21], indiscriminately erase all contextual details by zeroing irrelevant pixels—a method that, while reducing ambiguity, also discards valuable contextual cues. Recognizing the importance of these cues, we introduce purified masked matching (PMM), a delicate soft matching approach. PMM maintains the essential context by focusing on masked image features as the primary matching target. It employs a two-pronged strategy: reconstruction, where the image encoder works to replicate the masked feature, minimizing unnecessary semantics; and interaction, where the sketch encoder engages with the masked image feature via a transformer decoder, facilitating a purified cross-domain dialog. The objective of PMM is to refine the semantic correlation between sketches and images, enhancing the model's ability to differentiate among semantic categories (Fig. 1(d)).

In summary, our main contributions are fourfold:

1) We introduce PuXIM, which is specifically designed to address semantic ambiguity in sketch–image pairings. PuXIM enhances differentiation between sketches and images within the semantic space, illustrating the advantages of purified cross-domain matching in ZS-SBIR.
2) We design a visual cross-semantic (VxL) sampler to generate linguistic masks for nonessential regions by analyzing visual responses against semantic labels from natural images. This enables the acquisition of more accurate image representations to improve the matching process.
3) We propose PMM, which focuses on treating the masked features of natural images as the primary targets for matching. The reconstruction aspect minimizes unnecessary semantics, leading to a more refined visual representation, whereas the interaction component, which leverages a transformer decoder, deepens the understanding of cross-domain relationships between sketches and images.
4) We validate the effectiveness of PuXIM through extensive testing and comparison across several benchmarks, including Sketchy, TU-Berlin, and QuickDraw. Our results demonstrate that PuXIM sets new benchmarks, achieving state-of-the-art performance in ZS-SBIR.

II. Related Works

A. Zero-Shot SBIR

Given a query sketch, SBIR aims to retrieve category-specific photos from a gallery of multicategory photos [6], [22]. Recent SBIR frameworks leverage metric learning techniques, such as contrastive learning and triplet learning, to train relationships between sketch–image pairs. These methods ensure that representations of samples from the same category are close in the latent semantic space [16], [23], [24]. Koley et al. [25] further addressed the sketch abstraction problem by introducing an abstraction-aware SBIR framework. After impressive progress in the SBIR task, this framework naturally extended to the more challenging and practical zero-shot SBIR (ZS-SBIR) task. In ZS-SBIR, the goal is to generalize knowledge from training sketch–image pairs of partially seen categories to retrieve images of unseen categories in a zero-shot retrieval scenario. Furthermore, SAKE [5] employed additional semantic information from ImageNet [26] for knowledge distillation to assist the model in learning a shared semantic space for sketch–image pairs. Wang et al. [27] utilized the SAKE framework, which further introduces a memory bank for storing a large number of semantic features, thus solving the problem that semantic features can only be trained in mini-batches. Sketch3T [18] introduced two self-supervised learning methods, namely, contour sketches and stroke sequence learning, and proposed a framework trained during testing. TVT [28] trained a fusion model through knowledge distillation by using heterogeneous encoders for pre-trained sketches and images in a three-way framework. More recently, the cross-modal retrieval framework ZSE-SBIR was proposed to introduce cross-attention to solve the ZS-SBIR cross-modal matching problem and explore the interpretability of ZS-SBIR for the first time [16]. However, its cross-modal components in the retrieval task add a large computational burden to the inference process. Ren et al. [29] eliminated the sketch–image domain gap by generating large numbers of diverse photolike images. Sain et al. [30] proposed a CLIP-based [31] fine-tuning architecture for SBIR and achieved remarkable performance in the zero-shot scenario.

Sketches are composed of sparse strokes that succinctly represent a target and are devoid of extraneous background or semantic details. Existing zero-shot sketch-based image retrieval

(ZS-SBIR) methods match sketches to images but often fail to account for the rich background, semantic, and textural details inherent in natural images. This oversight causes models to focus excessively on ambiguous target regions within the images. In contrast, our proposed PuXIM leverages semantic labels from SBIR datasets and introduces the VxL sampler to mask ambiguous features, directing the model's attention to target regions corresponding to the sketch. Additionally, we design specialized reconstruction and interaction operations for image and sketch features. The reconstruction operation compels the image encoder to replicate masked features while preserving contextual information and emphasizing key regions. The interaction operation facilitates cross-modal learning by shuffling sketches and masked features, enhancing cross-domain interactions. PuXIM operates independently of specific model frameworks and can be seamlessly integrated into existing SBIR architectures, enabling the construction of a purified semantic space.

B. Vision and Language Pre-Training

Vision and language pretraining (VLP) aims to leverage the learning of both image and text features for application in downstream tasks. Numerous studies have addressed downstream multimodal tasks such as visual question answering (VQA) and natural language for visual reasoning (NLVR) [32], [33]. They primarily adhere to the following framework in their design: dual encoders [33], [34] have been employed to extract text features via methods such as BERT [35], and visual features have been extracted via CNN or ViT. A multimodal fusion module was utilized to compute cross-modal interactions. In terms of multimodal fusion, ALBEF [36] demonstrated the feasibility of momentum models in VLP and proposed aligning cross-modal features before fusion. ViLT [20] employed a minimalistic unified transformer architecture to receive inputs from the visual and textual sides simultaneously for multimodal alignment. Pixel-BERT [37] abandoned the approach of using expensive object detectors on the vision side by aligning pixel-level features with text features. These models all utilize a deep multimodal transformer to align the image–text representation. Instead, models such as CLIP [31] employed a relatively lightweight multimodal matching approach, calculating the similarity of image–text representations through contrastive learning to align multimodal features. Recent works have also investigated CLIP in explainable styles [38], [39], [40]. These VLP models typically require massive image–text pairs for training to acquire robust generalization capabilities for better transfer to downstream tasks, such as cross-modal image–text retrieval. Despite the success of these approaches for cross-modal retrieval, their contributions to the sketch community remain limited at present.

C. Masked Learning

Masked language modeling (MLM) was first introduced in natural language processing (NLP) [35], [41]. MLM randomly predicts masked sequences for upstream self-supervised learning and has achieved incredible success in downstream applications. Inspired by MLM, the computer vision community has also introduced masked image modeling (MIM) based on transformer-based architecture [42] and has made considerable progress in the field of self-supervised learning [43], [44], [45], [46]. BEiT [43], as a pioneer, first considered employing a masked image modeling task to train self-supervised visual encoders. MAE [44] embedded masked images via autoencoders and predicted missing pixels through lightweight decoders, thereby reconstructing the original image to obtain a powerful and generalizable visual encoder. Benefiting from the robust semantic modeling capability of MIM, it has also been introduced into unsupervised knowledge distillation frameworks to transfer knowledge from heavy models [47]. Unlike most of these works, which employ random masking strategies, we consider a more complex masking strategy for cross-domain sketch–image matching related to semantic descriptions.

III. Methodology

In this section, we detail each key module in our proposed framework, which consists of a VxL sampler, PMM reconstruction, and PMM interaction, as illustrated in Fig. 2.

A. Dual Encoders

Given a query sketch $S \in \mathbb{R}^{h\times w\times c}$ and a gallery image $I \in \mathbb{R}^{h\times w\times c}$, inspired by the success of the Vision Transformer (ViT) [48] in visual downstream tasks and cross-domain learning [49], we use a 12-layer ViT (*i.e.*, ViT-B/16) as the sketch encoder $f_\theta^S(\cdot)$ and image encoder $f_\theta^I(\cdot)$. An input sketch S or image I is embedded into the $p+1$ patch sequence: $\{P_{cls}, P_1, \ldots P_p\}$. We replace the [CLS] token P_{cls} with a retrieval token [RET] as a global visual representation. The information interaction is performed through multiple-head self-attention blocks to acquire global contextual information. Specifically, in the self-attention module, given the input X, we first obtain three matrices: the query (Q), key (K), and value (V) from the input by

$$Q = W_q X, K = W_k X, V = W_v X \tag{1}$$

where W_q, W_k and W_v are the corresponding weight matrices. The attention output is computed by

$$X_{\text{attn}} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V. \tag{2}$$

We employ dual encoders with shared parameters as the base architecture of PuXIM. During inference, cross-domain matching is performed by computing the cosine similarity between the sketch representation $p_S(S) \in \mathbb{R}^{1\times d}$ and the image representation $p_I(I) \in \mathbb{R}^{1\times d}$. Our key innovation is to obtain distraction-agnostic inference performance by learning purified sketch–image representations during the training phase.

B. Visual-Cross-Linguistic Sampler

To constrain the ambiguous semantics in images for purified sketch–image matching, we propose our VxL sampler to generate linguistic masks.

The sketch–image pairs used for training are mostly sourced from the internet. Weakly related images can be considered positive pairs, whereas positive samples related to the sketches can

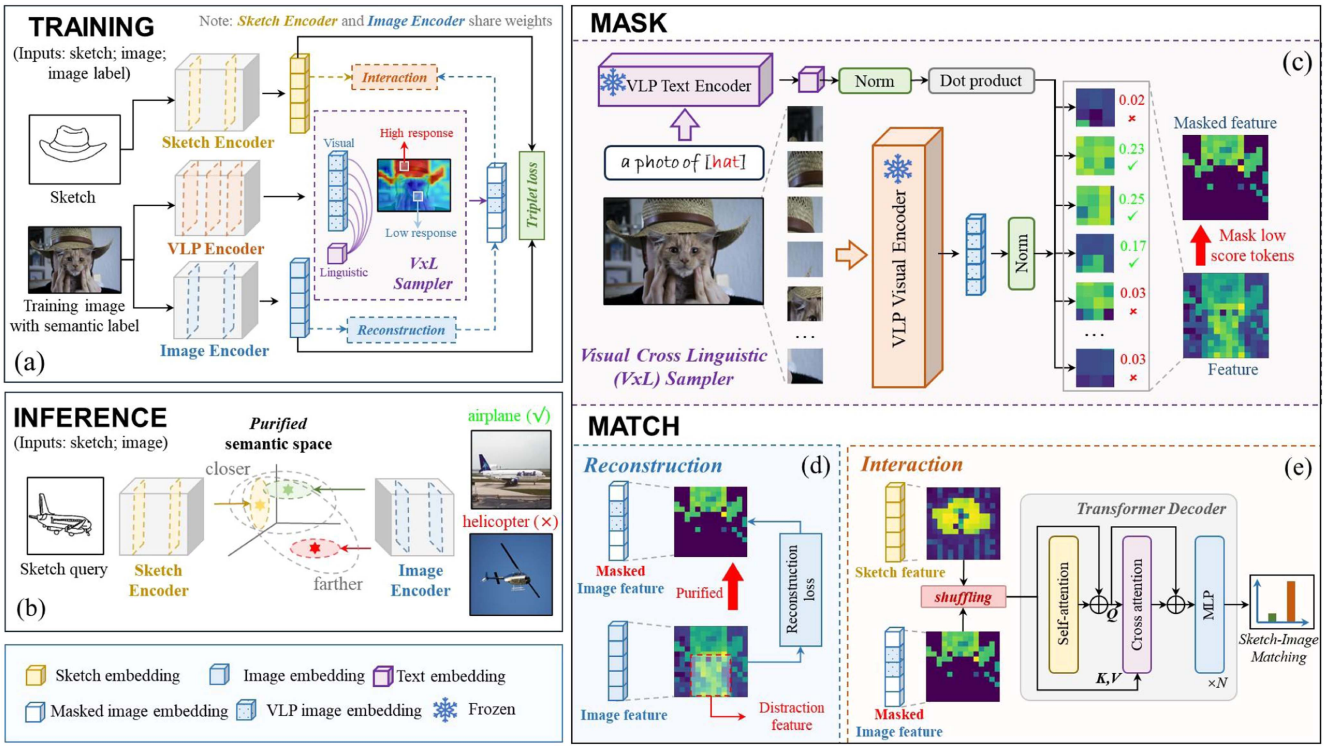

Fig. 2. Overview of the proposed PuXIM framework: (a) PuXIM training stage: complete structure, combining the VxL sampler with purified masked matching (reconstruction and interaction) for purified semantic space. (b) PuXIM inference stage: standard dual encoder setup without additional components for baseline evaluation in a purified semantic space. (c) VxL sampler's application of VLP for generating linguistic masks from semantic labels to filter out irrelevant features in natural images. (d) A reconstruction method that aligns original and masked image features to refine semantic accuracy. (e) Interaction introduces a transformer decoder to enhance sketch–image matching by processing both shuffled sketch and shuffled masked image patches for effective and fine-grained cross-domain interaction.

also be included in negative pairs. Matching such sketch–image pairs may easily lead to semantic ambiguity. To address this, we design a VxL sampler (Fig. 2(c)), which leverages the power of VLP to select essential tokens via freely available semantic labels in the training set $\mathcal{D}$. CLIP (contrastive language–image pretraining) [31], which is widely popular for open-set visual understanding tasks, consists of two separate encoders, one for images and one for text.

We obtain the manual labels for each image from the training set as the ground truth, whereas other semantic information present in the image is considered ambiguous semantics. CLIP ViT-B/16 is used as a visual encoder $f_\phi^I(\cdot)$ to embed images into visual features $v_{\text{CLIP}} \in \mathbb{R}^{(m+1)\times d}$ comprising m patches. Similarly, the text encoder $f_\phi^T(\cdot)$ embeds sentences of length s into text features $t_{\text{CLIP}} \in \mathbb{R}^{(s+1)\times d}$. We generate a text prompt "`a photo of [category]`" via semantic labels and calculate the similarity scores between the text T of "`[category]`" and each visual embedding token $f_\phi^I(I)_i \in \mathbb{R}^{m_i \times d}$ (m_i is the index of the i-th image patch). The score is represented as:

$$score = \frac{f_\phi^I(I)_i}{\left\| f_\phi^I(I)_i \right\|_2} \cdot \left(\frac{f_\phi^T(T)}{\left\| f_\phi^T(T) \right\|_2} \right)^\top \tag{3}$$

The top $(1-r) \times m$ proportion of the highest-scoring image tokens in VLP image feature $v_I(I) \in \mathbb{R}^{m\times d}$ are retained to obtain the masked image feature map $\widetilde{v}_I(I) \in \mathbb{R}^{m\times d}$ by VxL sampler, while the remaining tokens are masked out (r is the masking ratio). The VxL sampler that works during the training stage can be summarized as Algorithm 1.

Algorithm 1: Working Mechanism of the VxL Sampler on a Single Natural Image.

Input: A natural image I and its semantic label T from the training dataset $\mathcal{I}_s$.

1: Calculate image feature $f_\phi^I(I)$ through CLIP visual encoder $f_\phi^I(\cdot)$.
2: Calculate text feature $f_\phi^T(T)$ through CLIP text encoder $f_\phi^T(\cdot)$.
3: **for** $i = 1$ to m **do**
4: **repeat**
5: Calculate the similarity score between image token feature $f_\phi^I(I)_i$ with text feature $f_\phi^T(T)$ according to Equation (3).
6: **until** all image tokens are calculated
7: **end for**
8: Set the feature values of the lowest-scored image tokens to zero at a mask rate of r.

Output: masked image feature $\widetilde{v}_I$.

C. Purified Masked Matching

Designing a training approach that helps the model learn essential information within the region of interest is of utmost importance. A simplistic method to achieve this is to mask out irrelevant input tokens. However, the background environment of an image contains certain semantic priors that help the model better differentiate specific categories in the image. Removing this portion of information also results in the loss of visual information. Therefore, we propose PMM to manipulate image representation at the feature level. PMM treats $\widetilde{p}_I$ as the matching target, on which the reconstruction and interaction operations are applied to the image encoder and sketch encoder, respectively.

1) Reconstruction: Our PMM reconstruction is simple, *i.e.*, the image encoder is encouraged to output purified representations to suppress redundant semantics while preserving the original contextual information. In contrast to other mask modeling methods, PMM reconstruction regards the masked image features $\widetilde{v}_I$ as the reconstruction target, aligning the image encoder output p_I with $\widetilde{v}_I$ as much as possible, as shown in Fig. 2(d). Our reconstruction loss computes the mean square error (MSE) loss between the original and masked image features, which is specifically represented as:

$$L_{rec}(p_I, \widetilde{v}_I) = \frac{1}{\mathcal{B}} \sum_{b=1}^{\mathcal{B}} \sum_{l=1}^{L} \left(\left\| p_I^l - \widetilde{v}_I^l \right\|_2^2 \right) \tag{4}$$

where L is the sum of the reconstruction layers. $\mathcal{B}$ is the mini-batch size.

Moreover, part of the image subset has a large amount of redundant semantics, and the feature signals of the ground truths are weak; hence, we compute the mean and standard deviation of the remaining tokens to obtain normalized mask features and ensure that the retained features have sufficient feature responses to improve the reconstruction quality.

2) Interaction: To further explore the cross-domain relationships between sketches and images in semantic space, we propose an interaction operation that consists of transformer decoders for cross-domain matching, as shown in Fig. 2(e). Specifically, our cross-domain decoder (CDD) contains self-attention, cross-attention, and MLP modules. For cross-attention, the tokens from one domain are considered queries (Q), and the tokens from the other domain are considered Keys and Values (K, V).

Our proposed PMM interaction leverages masked image features as interaction targets to determine whether the sketch and image share the same semantic category. To enhance this process, we introduce a shuffling operation that randomizes patches between the image and the sketch. Unlike globally ordered patch matching, shuffling requires the model to evaluate semantic relevance on the basis of local relations in unordered patches, fostering the learning of finer-grained interactions. This approach is particularly effective in scenarios with high semantic ambiguity, where achieving such precision without semantic purification would be challenging. In our implementation, the shuffled output from the sketch encoder serves as Q_S, and the shuffled image feature $\widetilde{v}_I$ serves as K_I^{msk} and V_I^{msk}. The cross-attention is calculated by

$$X_{\text{cross}} = \text{softmax}\left(\frac{Q_S \left(K_I^{msk}\right)^\top}{\sqrt{d}} \right) V_I^{msk}. \tag{5}$$

Our interaction loss is a binary cross-entropy loss that is optimized to classify whether a sketch–image pair matches. The loss can be written as follows:

$$\begin{aligned} L_{int} = \mathbb{E}_{\left(p_S^A, \widetilde{v}_I^A\right) \sim \mathcal{D}} [& y \log \left(\text{CDD}\left(p_S^A, \widetilde{v}_I^A\right) \right) \\ & + (1-y) \log \left(1 - \text{CDD}\left(p_S^A, \widetilde{v}_I^A\right) \right)] \end{aligned} \tag{6}$$

where y is the label (*, i.e.,* 1 for the matching pair and 0 otherwise), and CDD is a cross-domain decoder block.

The interaction operation involves two key issues: (1) We observe that the matching task is challenging to optimize because of the high heterogeneity and semantic ambiguity between sketches and images. Therefore, we utilize masked features as matching targets for purified cross-modal interactions. (2) We introduce an additional semantic space through a decoder during the training phase to explore cross-domain relationships between sketches and images while avoiding the use of computationally expensive cross-attention during the inference phase, as shown in Fig. 2(b).

D. Total Loss

We exploit three losses to train our framework: a triplet loss applied on the retrieval token, a reconstruction loss, and an interaction loss on our PMM. The triplet loss can be written as:

$$L_{tri} = \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \max \left\{ d\left(x_a, x_p\right) + \alpha - d\left(x_a, x_n\right), 0 \right\} \tag{7}$$

where $d(x_i, x_j) = ||x_i - x_j||_2$, α is the margin, and x_a, x_p, and x_n are the anchor sketch, positive and negative photos, respectively. To this end, the overall loss is summed as:

$$L = L_{tri} + \lambda_{rec} L_{rec} + \lambda_{int} L_{int} \tag{8}$$

where λ_{rec} and λ_{int} are the corresponding loss weights.

IV. Experiments

A. Datasets and Settings

1) Datasets: We evaluate our method using four popular datasets for category-level ZS-SBIR. TU-Berlin [15] contains 250 categories, with 80 sketches in each, extended to 204,489 images. Following [50], we split it into 30 classes for testing and 220 classes for training. Sketchy [14] contains 75,471 sketches in 125 categories with 100 images each. The extended version [51] has an extra 60,502 images from ImageNet [26]. Following [12], we split it into 104/100 classes for training and 21/25 classes for testing for the zero-shot setup. The QuickDraw Extended [52] full version contains over 50 million sketches in 345 categories. After expanding it with images, ZS-SBIR introduces a subset of 110 categories with 330,000 sketches and 204,000 photos. The statistical results of the ZS-SBIR datasets are listed in Table I. For FG-ZS-SBIR, we use 104 classes from the Sketchy dataset for training and 21 classes for zero-shot testing.

TABLE I
STATISTICS ON THE NUMBER OF LABELS PER CATEGORY AND THE NUMBER OF SKETCHES AND IMAGES OF THE ZS-SBIR DATASETS

Dataset	Seen classes			Unseen classes		
	Num	Sketch	Image	Num	Sketch	Image
TU-Berlin	220	15400	176081	30	2400	27989
Sketchy	100	55252	68401	25	15229	17101
Sketchy Split	104	57587	72949	21	12694	12553
QuickDraw	80	240080	149433	30	92991	54151

2) Implementation Details: We implement our method in PyTorch on a 24 GB Nvidia RTX 4090 GPU. For sketch and image inputs, we use two shared ViT-B/16 [48] pre-trained on ImageNet [26] as encoders. For the VxL sampler, we use CLIP [31] with ViT-B/16 as the visual encoder. The input image size is set as 224×224 with margin parameter $\mu = 2.0$, and prompts are trained via the Adam optimizer with a learning rate $1e{-}5$ for 40 epochs and batch size 64. VLP is fine-tuned on the corresponding dataset for better mask generation. The last three layers ($L = 3$) of features from the image encoder and VLP encoder are extracted for reconstruction, whereas the last layer feature is extracted for interaction. For interaction, we use a transformer decoder (without masking) with 4 layers and 4 heads. All dimensions of the embeddings in the latent space are 768. The values of λ_{rec} and λ_{int} are empirically set to 0.05 and 0.1, respectively..

3) Evaluation Protocol: According to recent SBIR methods [16], [28], the mean average precision (mAP) and precision of the top 100 (Prec@100) and top 200 (Prec@200) methods are reported following the standard evaluation protocol. For the FG-ZS-SBIR, following Sain et al. [30], accuracy is measured within a single category, *i.e.*, acc.@q, reflecting the percentage of images in the top-q list that match the input sketch. We report acc.@1 and acc.@5.

B. Comparison With State-of-The-Art ZS-SBIR

We compare the ZS-SBIR frameworks initialized on ImageNet [26] across the TU-Berlin Ext, Sketchy Ext, Sketchy Ext Split, and QuickDraw datasets. The evaluated frameworks include SEM-PCYC [53], SAKE [5], PDFD [55], PCMS [56], and AMA [59]. They all use a common CNN to extract sketch–image features and introduce text semantic features. SketchGCN [54] introduces a GCN model for semantic preservation. DSN [17] incorporates a memory bank into a contrastive learning framework. TCN [63], similar to the SAKE [5] framework, introduces additional semantic labels from ImageNet for knowledge distillation. CA [27] introduces a memory bank, which addresses the problem that SAKE can only learn a limited number of semantic signals in a mini-batch. More recently, Sketch3T [18], TVT [28], CIIM [60], and ZSE-SBIR [16] have used ViT as the encoder for sketch–image representations. They employ additional self-supervised tasks, knowledge distillation, cross-modal similarity learning, and cross-attention encoders to reduce the semantic gap between sketches and images. We also compare recent CLIP-based frameworks, including CLIP-AT [30], TLT [64], and CLIP-MA [62]. We employ the same training strategy and model as Sain et al. [30] for ZS-SBIR testing, with the additional inclusion of PMM reconstruction and interaction. The results are listed in Table II .

As shown in Table II, even though our framework uses the common dual encoder design during the inference stage, it consistently outperforms the SOTA methods. In particular, it achieves an absolute improvement of 12.5% in mAP@200 on the challenging Sketchy Ext Split dataset. Additionally, on TU-Berlin and Sketchy Ext, PuXIM achieves absolute improvements of 0.5% and 1.2% in terms of mAP, respectively, although it performs slightly worse than some methods on QuickDraw. Moreover, PuXIM (CLIP) achieves state-of-the-art performance on four datasets, outperforming existing CLIP-based methods. Notably, on the Sketchy Ext dataset, it achieves a significant absolute improvement of 5% in mAP. Importantly, PuXIM operates solely as a purification strategy during training, introducing no additional computational cost and not relying on the CLIP text encoder during inference. This makes it more flexible and efficient than fully CLIP-based models.

Moreover, we compare our method to a similar architecture with ViT-B/16 as the backbone ZSE-SBIR [16] by selecting certain sketch queries to show qualitative results on Sketchy in Fig. 3. For more realistic style sketches, both PuXIM and ZSE-SBIR retrieve relevant images well, except for "windmill", where ZSE-SBIR retrieves "bell". For creative style sketches, PuXIM retrieves the correct "parrot" at the top 1, and some failure cases are reasonable. However, ZSE-SBIR ignores fine-grained strokes in sketches and provides the wrong top 5 retrieves. Although both methods incorporate masking, the ZSE-SBIR mask primarily reduces computational complexity and lacks directionality, often resulting in a focus on explicit foreground features. This leads to increased semantic ambiguity, with the model relying on shape matching rather than semantic alignment. In contrast, PuXIM generates precise masks via semantic labels, enabling it to retain critical target regions while reconstructing ambiguous features. This approach ensures robust semantic alignment and effective cross-domain interaction within a purified semantic space.

C. Comparison With State-of-The-Art FG-ZS-SBIR

Category-level ZS-SBIR considers images across all categories as equivalent, focusing only on semantic relationships while neglecting fine-grained aspects such as texture, pose, and orientation. FG-ZS-SBIR is more challenging, requiring the model to pay attention to fine-grained details alongside semantic alignment.

For fair comparison, we evaluate models initialized on ImageNet [26], including CrossGrad [65], CC-DG [66], and ZSE-SBIR [16]. PuXIM outperforms the state-of-the-art ZSE-SBIR by 4.05% in acc.@1. While both methods use masking mechanisms, ZSE-SBIR primarily preserves foreground targets, which may lead to errors in high semantic ambiguity images. PuXIM leverages text labels to focus directly on targets matching the sketch, achieving precise semantic alignment and superior performance in FG-ZS-SBIR.

The visualization results in Fig. 3 highlight PuXIM's effectiveness. It accurately retrieves corresponding images in the

TABLE II
QUANTITATIVE COMPARISON OF PUXIM AGAINST EXISTING FRAMEWORKS ON POPULAR ZS-SBIR DATASETS. "-": NOT REPORTED, "INIT": MODEL PARAMETERS INITIALIZATION, "∗": OUR REPRODUCTION.

INIT	Method	TU-Berlin Ext		Sketchy Ext		Sketchy Ext Split		QuickDraw Ext	
		mAP	Prec@100	mAP	Prec@100	mAP@200	Prec@200	mAP	Prec@200
ImageNet	SEM-PCYC [15]	0.297	0.426	0.349	0.463	-	-	-	-
	SAKE [37]	0.475	0.599	0.547	0.692	0.497	0.598	0.130	0.179
	SketchGCN [75]	0.324	0.505	0.382	0.538	-	-	-	-
	StyleGuide [16]	0.254	0.355	0.376	0.484	0.358	0.400	-	-
	PDFD [64]	0.483	0.600	0.661	0.781	-	-	-	-
	PCMS [11]	0.424	0.517	0.523	0.616	-	-	-	-
	DSN [61]	0.484	0.591	0.583	0.704	-	-	-	-
	BDA-SketchRet [5]	0.375	0.504	0.437	0.514	0.556	0.458	0.154	0.355
	SBTKNet [57]	0.480	0.608	0.553	0.698	0.502	0.596	-	-
	Sketch3T [50]	0.507	-	0.575	-	-	-	-	-
	TVT [55]	0.484	0.662	0.648	0.796	0.531	0.618	0.149	0.293
	AMA [71]	0.429	0.592	0.491	0.585	0.548	0.684	0.112	0.192
	CA [60]	0.542	0.622	0.633	0.709	-	-	-	-
	ACNet [46]	0.577	0.658	-	-	0.517	0.608	-	-
	CIIM [66]	0.569	0.693	0.722	0.818	0.541	0.644	0.167	0.305
	ZSE-SBIR [34]	0.542	0.657	0.698	0.797	0.525	0.624	0.145	0.216
	Ours (ViT-B/16)	**0.582**	**0.673**	**0.734**	**0.810**	**0.681**	**0.633**	**0.150**	**0.255**
CLIP	CLIP-AT [48]	0.651	0.732	-	-	0.723	0.725	0.202	0.388
	TLT [73]	0.615	0.695	0.779	0.843	0.661	0.730	-	-
	CLIP-MA* [39]	0.668	0.741	0.753	0.797	0.694	0.708	0.229	0.405
	Ours (CLIP)	**0.697**	**0.750**	**0.829**	**0.869**	**0.746**	**0.799**	**0.235**	**0.402**

Fig. 3. Top-5 ZS-SBIR (top) and FG-ZS-SBIR (bottom) results obtained via PuXIM and ZSE-SBIR on the Sketchy Ext and Sketchy Ext Splits. The green ticks denote correctly retrieved candidates, and the red crosses indicate incorrect retrievals. PuXIM has an advantage when dealing with more creative sketches. Specifically, when given a sketch of a "parrot", its top 1 retrieval result is correct, whereas ZSE-SBIR provides incorrect images in all of its top 5 retrieval results. This finding demonstrates that purified masked matching can help the model learn a deeper sketch–image cross-domain relationship, thereby establishing a more accurate semantic space.

TABLE III
QUANTITATIVE COMPARISON OF PUXIM AGAINST EXISTING FG-ZS-SBIR FRAMEWORKS ON SKETCHY DATASETS. "INIT": MODEL PARAMETERS INITIALIZATION.

INIT	Method	acc@1	acc@10
ImageNet	CrossGrad [53]	13.40	34.90
	CC-DG [43]	22.60	49.00
	ZSE-SBIR [34]	23.97	49.52
	PuXIM (ViT)	**28.02**	**57.82**
CLIP	CLIP-AT [48]	28.68	62.34
	CLIP-MA [39]	29.96	58.53
	PuXIM (CLIP)	**31.83**	**64.63**

TABLE IV
COMPARISON OF THE EFFECTS OF DIFFERENT VXL SAMPLER MASKING RATES AND THE APPLICATION OF RECONSTRUCTION LOSS ON PUXIM-FULL FOR SKETCHES ON SKETCHY EXT DATASET. *SK* REPRESENTS THE COMPONENT USED ON THE SKETCH SIDE.

Architecture	Rate (sk)	VxL (sk)		VxL (sk)+$L_{rec}(sk)$	
		mAP	Prec@100	mAP	Prec@100
Ours-full	0	0.734	0.810	0.734	0.810
	0.3	0.626	0.747	0.678	0.780
	0.5	0.527	0.650	0.651	0.755
	0.7	0.463	0.531	0.613	0.701

TABLE V
COMPARISON OF PUXIM-FULL VIA DIFFERENT VXL SAMPLER MASKING RATIOS ON SKETCHY AND TU-BERLIN DATASETS

Architecture	Rate	TU-Berlin Ext		Sketchy Ext	
		mAP	Prec@100	mAP	Prec@100
Baseline	0	0.450	0.588	0.579	0.718
Ours-full	0.3	0.517	0.641	0.679	0.767
	0.5	0.543	0.656	0.688	0.776
	0.7	**0.582**	**0.673**	**0.734**	**0.810**
	0.8	0.561	0.672	0.710	0.782
	0.9	0.539	0.661	0.672	0.746

top-5 list. While plausible errors occur (*e.g.*, a "tree" retrieved for a "windmill" sketch), ZSE-SBIR often retrieves incorrect categories owing to its reliance on contour-based matching, such as retrieving a "rhinoceros" for a "cow" sketch, ignoring fine-grained details such as body texture.

In CLIP-based models, CLIP-AT and CLIP-MA integrate text features but are limited in distinguishing cross-class semantics, making precise matching challenging. PuXIM enhances textual guidance for semantic purification, achieving SOTA performance in both acc. @1 and acc. @5 for FG-ZS-SBIR.

D. Quantitative Analysis

To provide a more reliable quantitative analysis, all our subsequent experiments use ViT-B/16 pre-trained on ImageNet [26] as the backbone of PuXIM.

1) Masking Ratio of VxL Sampler: The VxL sampler relies on an appropriate mask ratio; a too-low ratio fails to filter out redundant semantics, whereas a too-high ratio damages the original object semantics. We test our full PuXIM performance with different masking ratios. The results are listed in Table V. We conclude that the model performs optimally with a masking rate of 0.7. To obtain a qualitative sense of our linguistic masks, see Fig. 4. With a high masking ratio of 0.7, the retained tokens can preserve information at key locations and eliminate ambiguous semantics effectively, *e.g.*,, "cat" and "book" and "eyeglasses" and "cat". The results for different masking ratios are consistent with our expectations, with the lowest performance observed at rates of 0.3 and 0.9.

Fig. 5 shows the visualization results of different mask ratios. For most natural images, 70% can be adapted, and the size of the objects in the images does not significantly affect the results. Smaller objects can be extracted precisely, whereas larger objects do not suffer from excessive semantic loss.

To investigate the effect of sketch masking, we apply two masking strategies within the PuXIM framework: (1) directly erasing low-response patches via the VxL sampler and (2) applying the proposed PMM reconstruction loss to the sketch features. The visualization of masked sketches is shown in Fig. 5.

The performance results are summarized in Table IV. Masking sketches results in performance degradation. This is likely due to the limited generalizability of VLP models such as CLIP to sketch data, especially abstract sketches, where locating target features via semantic labels is challenging. Masking disrupts the sparse visual cues that are critical for sketch representation. The performance is better at lower masking rates, where less information is removed.

Furthermore, applying the proposed PMM reconstruction loss mitigates the performance drop, as it preserves the original sketch features. At a high masking rate ($r = 0.7$), PuXIM with sketch mask reconstruction outperforms the naive hard mask strategy by 15% in mAP, demonstrating the effectiveness of reconstruction in maintaining sketch integrity.

TABLE VI
COMPARISON OF PMM RECONSTRUCTION VIA DIFFERENT VISUAL ENCODERS ON SKETCHY EXT DATASET. "$*$" REPRESENTS SELF-MASKING.

Encoder	L_{tri}		$L_{tri} + L_{rec}$	
	mAP	Prec@100	mAP	Prec@100
RN18 [20]	0.383	0.548	0.440	0.585
RN50 [20]	0.452	0.586	0.511	0.657
DINO-S/8 [72]	0.527	0.639	0.615	0.724
ViT-B/16* [58]	0.579	0.715	0.653	0.774
ViT-B/16 [58]	**0.579**	**0.715**	**0.680**	**0.788**

2) PMM Reconstruction: To validate the generalizability of PMM reconstruction, we exclude the PMM interaction and adopt several common models as sketch–image encoders for validation on Sketchy Ext, including the ViT [48], DINO [67] and ResNet [68] series. Moreover, given the extensive training of the VLP visual encoder on upstream data, its feature map holds richer priors. To maintain rigor, we also test a self-mask strategy on ViT-B/16 output features to ensure consistency between image features and masked image features, thereby mitigating potential benefits from the powerful semantic richness associated with masked VLP features. As shown in Table VI, reconstruction is applied to most common models, and lightweight ResNet-18 and DINO-S/8 achieve significant improvements of 5.7% and 8.8% in mAP, respectively. Our standard architecture ViT-B/16 improves the mAP by 10.1%. Notably, compared with the VLP mask feature in our standard PuXIM, the performance of the ViT self-masking feature slightly decreases (2.7% mAP). This is expected, as the rich semantic features in CLIP transfer

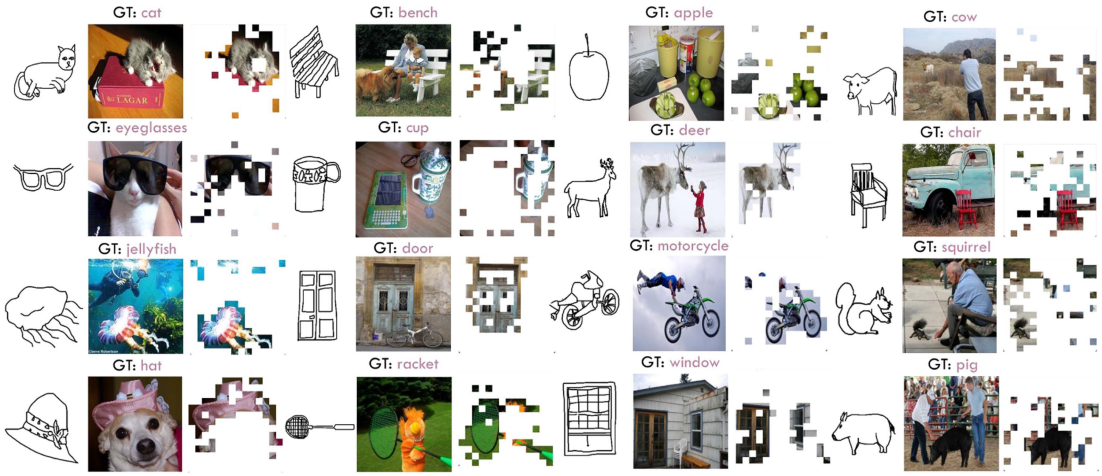

Fig. 4. Ambiguous semantic sketch–image pairs and their masked tokens ($r = 0.7$). Semantic labels capture visual concepts for purified matching. When the masking threshold is 0.7, it captures most of the effective feature regions while eliminating redundant semantics (*e.g.*,, "book", "dog", "cat").

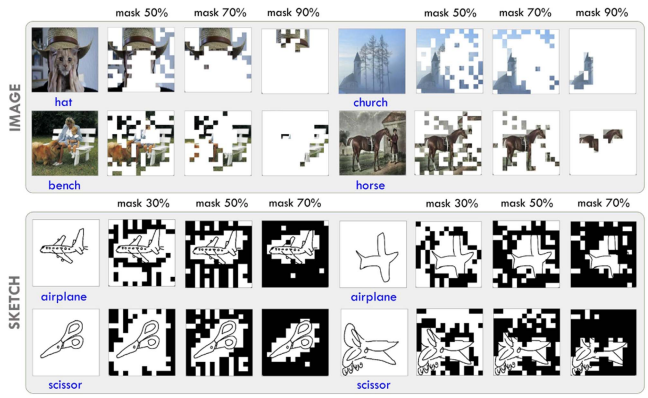

Fig. 5. Visualization examples for different mask rates on the image (top) and sketch (bottom). For the image side, a low mask rate may introduce ambiguous semantics, and a high mask rate will destroy the real semantic target, but the performance is higher than that of the *baseline*, showing that the method can generalize, with the best performance in both datasets at $r = 0.7$. For the sketch side, since sketches do not suffer from semantic ambiguity, the best performance is achieved when the mask rate is set to 0. For abstract sketches, a high masking ratio in the VxL sampler may disrupt the original sketch information.

TABLE VII
SOTA METHODS USING PMM RECONSTRUCTION ON SKETCHY EXT DATASET

Method	original		purified	
	mAP	Prec@100	mAP	Prec@100
SAKE [37]	0.551	0.687	0.624	0.713
TVT [55]	0.633	0.741	0.671	0.790
ZSE-SBIR [34]	0.684	0.778	0.717	0.813

TABLE VIII
COMPARISON OF PMM INTERACTION VIA DIFFERENT CROSS-DOMAIN INPUTS. p_I: OUTPUT FROM IMAGE ENCODER, p_S: OUTPUT FROM SKETCH ENCODER, v_I: OUTPUT FROM VLP VISUAL ENCODER $\widetilde{p_I}$: MASKED OUTPUT FROM IMAGE ENCODER, $\widetilde{v_I}$: MASKED OUTPUT FROM VLP VISUAL ENCODER.

int.	**TU-Berlin Ext**		**Sketchy Ext**	
	mAP	Prec@100	mAP	Prec@100
w/o int.	0.450	0.588	0.579	0.718
p_I, p_S	0.449	0.585	0.573	0.693
v_I, p_S	0.468	0.584	0.599	0.717
$\widetilde{p_I}$, p_S	0.499	0.627	0.637	0.732
$\widetilde{v_I}$, p_S	**0.512**	**0.615**	**0.656**	**0.748**

to the base model during the reconstruction process lead to some improvements. This process is similar to knowledge distillation. As expected, most of the performance improvement comes from the masking effect (7.4% mAP), indicating the effectiveness of semantic purification.

In addition to being applied in different base triplet frameworks, PMM reconstruction can also be easily embedded into any existing architecture as an additional component during the training stage. We retest several existing representative SOTA architectures and evaluate the benefits of PMM reconstruction. As shown in Table VII, PMM reconstruction effectively improves the performance of existing methods, with a maximum mAP improvement of 7.3% on SAKE [5], and the current SOTA method, ZSE-SBIR [16], also gains a 3.3% mAP improvement. This demonstrates its ability to generalize across other architectures, and our proposed approach can assist any future architecture in achieving better cross-domain matching.

3) PMM Interaction: To further verify the validity of the PMM interaction individually, we remove the reconstruction operation and emphasize comparing the results of the interaction without our linguistic mask. Specifically, we employ different combinations of feature interactions, including the image features p_I and sketch features p_S from the dual encoders, the VLP image features v_I, the masked image features $\widetilde{p}_I$, and the masked VLP image features $\widetilde{v}_I$. Table VIII shows that previous works employing direct cross-domain interactions without masks perform even worse than the baseline in terms of the mAP (0.449 vs. 0.450 on TU-Berlin, 0.573 vs. 0.579 on Sketchy). Owing to the rich prior semantic features of VLP image features, there is a slight improvement when interacting with sketch features. Furthermore, the proposed purified masked interaction benefits from distraction-agnostic pixel-level correspondence, leading to

TABLE IX
COMPARISON OF DIFFERENT NUMBERS OF CROSS-DOMAIN DECODER LAYERS OF PMM INTERACTION

num.	TU-Berlin Ext		Sketchy Ext	
	mAP	Prec@100	mAP	Prec@100
2	0.493	0.609	0.620	0.728
4	0.512	0.615	0.656	0.748
6	0.464	0.587	0.598	0.732

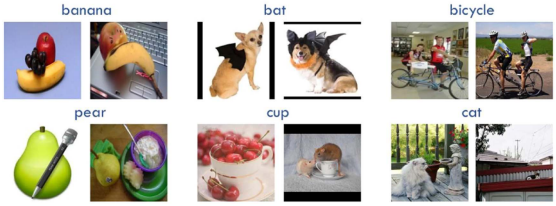

Fig. 6. Examples of the high semantic ambiguity training dataset. Each natural image contains the semantics of several non-ground-truth objects. The models train under conditions of semantic ambiguity, thereby posing a greater challenge for the zero-shot scenario.

significant improvements in both feature interaction schemes. Compared with the baseline, the standard PMM achieves significant improvements in the mAPs of 6.2% and 7.7% on TU-Berlin and Sketchy, respectively. The results indicate that performing cross-domain interaction under conditions of high semantic ambiguity is risky, as it can further distract the model. However, by applying the proposed PMM interaction to purify natural images, precise cross-domain interaction can be achieved, enabling the model to learn effective pixel feature correspondences between sketches and images.

We also investigate how the complexity of interaction affects the performance in Table IX, where we train our model with different numbers of layers of cross-domain decoders. The model performs better when there are more layers in the decoders (2 vs. 4 layers). However, excessively heavy decoders (6 layers) can also degrade performance. We argue that this is because the increased parameters of the decoders cause the model to focus excessively on interactions, thereby weakening the influence of the reconstruction components, resulting in the high semantic ambiguity condition of the cross-domain interactions mentioned.

4) High Semantic Ambiguity ZS-SBIR: We customize a training set focused on semantic ambiguity conditions to further validate the model's performance. The training set contains 50 categories (*e.g.*, banana, bicycle, cup) with a total of 934 highly semantically ambiguous images and 34532 sketches. The sketches and images are selected from the training set of the Sketchy Ext dataset, and each natural image contains several other objects or has a complex background texture in addition to the ground truth. The zero-shot validation set is consistent with Sketchy Ext. Some examples are shown in Fig. 6. We choose several popular frameworks to test the performance, including SAKE [5], ZSE-SBIR [16], CLIP [30], and our proposed PuXIM.

Table X shows that PuXIM achieves an mAP of 0.239 on just 934 highly semantically ambiguous natural images, surpassing all SOTA methods. Notably, SAKE, which also uses semantic

TABLE X
QUANTITATIVE COMPARISON OF PUXIM AGAINST EXISTING FRAMEWORKS ON HIGH SEMANTIC AMBIGUITY DATASET. THE SEMANTIC COLUMN REPRESENTS WHETHER THE METHOD USES SEMANTIC LABEL INFORMATION.

Methods	Backbone	Semantic	mAP	Prec@100
SAKE [37]	CSE-Res50	✓	0.144	0.175
SAKE [37]	ViT-B/16	✓	0.179	0.152
ZSE-SBIR [34]	ViT-B/16	✗	0.135	0.192
CLIP [48]	CLIP ViT-B/16	✗	0.149	0.144
PuXIM (Ours)	ViT-B/16	✓	**0.239**	**0.257**

labels, outperforms other methods, particularly after employing ViT-B/16, even exceeding CLIP (mAP of 0.179 vs. 0.149). We argue that SAKE potentially addresses the issue of semantic ambiguity by preserving knowledge from semantic labels. A part of the model's benefits comes precisely from this. The proposed masking strategy provides more effective semantic purification, surpassing SAKE (ViT-B/16) by 6% in terms of mAP. In contrast, ZSE-SBIR and CLIP, which do not use semantic information, perform similarly and fall behind methods that use semantic information. This suggests that severe model distraction poses a significant challenge in zero-shot scenarios.

5) ZS-SBIRs With Different Sketch Qualities: We evaluate the performance of our method in zero-shot scenarios across varying sketch quality levels. To quantify sketch quality, Yang et al. [69] proposed the geometric aware classification layer, which replaces the dense/softmax combination to predict a quality score between 0 and 1 through classification. However, as this method relies on stroke sequences and lacks subjective human evaluation, we adopt a more concise approach.

The Sketchy dataset [14] provides five quality labels: (1) correct, (2) contains environment details or shading, (3) incorrect pose or perspective, (4) ambiguous, and (5) erroneous. Ignoring "incorrect pose or perspective" and "contains environment details or shading," we redefine "correct" as good (1,137 sketches), "ambiguous" as medium (1,137 sketches), and "erroneous" as bad (917 sketches). Using these labels, we construct a dataset for quality prediction and train MobileNetV3 [70], following the implementation details of Liu et al. [71], achieving a final recognition accuracy of 90.85% (with 10% of sketches for validation).

The trained model predicts sketch quality across three SBIR datasets: Sketchy Ext, TU-Berlin, and QuickDraw. As shown in Fig. 7, Sketchy has the highest proportion of high-quality sketches, TU-Berlin contains more medium-quality sketches, and QuickDraw features the most low-quality sketches. We train the ZS-SBIR methods on Sketchy Ext and TU-Berlin Ext and validate them on all three datasets. This setup reflects a realistic and challenging zero-shot scenario, where the sketch styles in the training and testing sets differ significantly.

As shown in Table XI, the model's performance improves with the quality of the sketches across different datasets. PuXIM consistently performs the best under varying sketch qualities. For low-quality, rough sketches, PuXIM demonstrates better adaptability, and due to a more accurate semantic space and fine-grained matching, PuXIM shows significant improvement with high-quality sketches. Particularly in the TU-Berlin dataset, the mAP for high-quality sketches reaches 0.622, which is 11%

TABLE XI
COMPARISON OF DIFFERENT SKETCH QUALITIES."QS", "TS" AND "SS" DENOTE QUICKDRAW-STYLE, TU-BERLIN-STYLE, AND SKETCHY-STYLE SKETCH QUALITY. THE BACKBONE OF ALL COMPARATIVE METHODS IS VIT-B/16 FOR FAIR COMPARISONS.

Method	Sketchy Ext						TU-Berlin Ext					
	QS		TS		SS		QS		TS		SS	
	mAP	Prec@100	mAP	Prec@100	mAP	Prec@100	mAP	Prec@100	mAP	Prec@100	mAP	Prec@100
SAKE	0.194	0.268	0.394	0.517	0.607	0.741	0.230	0.294	0.511	0.624	0.488	0.594
ZSE-SBIR	0.231	0.345	0.466	0.562	0.684	0.778	0.262	0.357	0.544	0.660	0.512	0.625
PuXIM	0.263	0.360	0.493	0.585	0.734	0.810	0.279	0.393	0.582	0.673	0.622	0.677

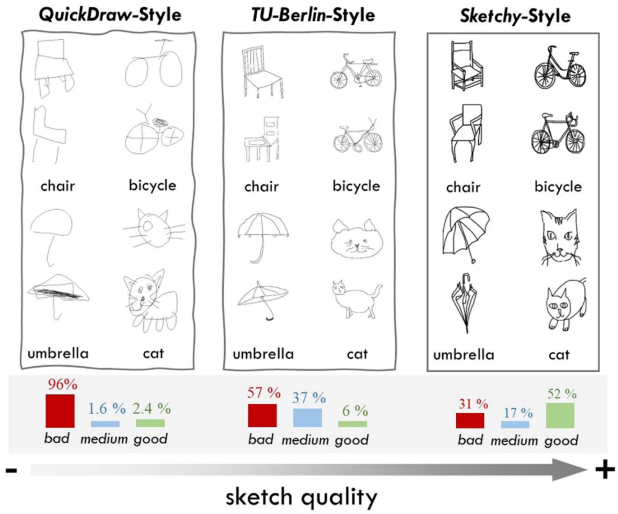

Fig. 7. Sketch datasets with different sketch qualities. QuickDraw sketch creators do not have any professional drawing skills, and the drawing time is limited to 20 seconds, resulting in most sketches being rough and abstract. TU-Berlin sketch creators are given 30 minutes to draw, with a potential list of images provided as prompts, leading to higher sketch quality. The goal of Sketchy is to create fine-grained and instance-level associations with natural images. Sketch creators must base their sketches on specific photos to ensure sufficient similarity, with no time limit on drawing, resulting in the highest sketch quality.

TABLE XII
COMPARISON OF DIFFERENT MASK STRATEGY ON SKETCHY EXT. VXL SAMPLER[*ORI.*] REPRESENTS THE RESULT OF DIRECTLY ERASING THE AMBIGUOUS SEMANTICS ON THE ORIGINAL IMAGE, AND VXL SAMPLER[*FEA.*] REPRESENTS THE RESULT OF ERASING THE AMBIGUOUS SEMANTICS ON THE FEATURE MAP AND INCORPORATING THE PMM.

Mask strategy	Methods	mAP	Prec@100
hard	VxL sampler[*ori.*]	0.526	0.679
	Grounding-DINO [38]	0.553	0.692
	Grounded SAM [47]	0.565	0.703
soft	VxL sampler[*fea.*]	0.734	0.810

higher than that of ZSE-SBIR, confirming its ability to generalize across sketches of different quality levels.

6) Comparison of Different Masking Strategies: We evaluate hard mask strategies, which directly erase ambiguous pixels, under the ZS-SBIR settings using the VxL sampler, Grounding-DINO [72], and Grounded-SAM [73]. Grounding-DINO locates targets via detection boxes, whereas Grounded-SAM applies the Segment Anything Model (SAM) [74] for semantic segmentation within these boxes, as shown in Fig. 8. The results in Table XII show that while these methods perform well in masking, the VxL sampler's soft mask strategy offers distinct advantages. Unlike hard masks, the VxL sampler computes patch-text

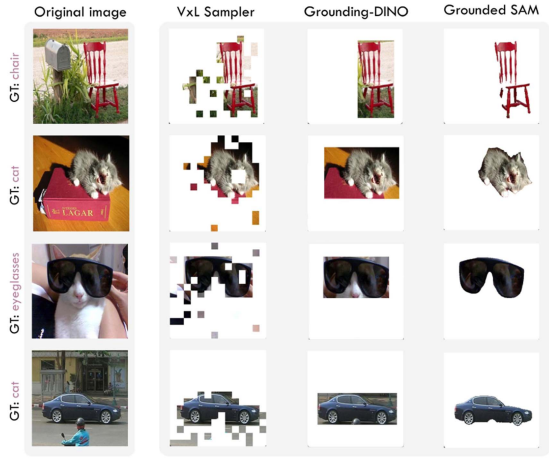

Fig. 8. Visualization of different masking strategies. The confidence threshold for the bounding box in Grounding-DINO is 0.3, and the matching threshold between the text prompt and the bounding box is 0.25.

similarities and operates at the feature map level via the PMM, allowing the model to retain contextual information while focusing on target regions. Hard masks, in contrast, disrupt image semantics by erasing ambiguous pixels and fail to generalize in zero-shot scenarios where semantic labels are unavailable. The soft mask strategy with PMM outperforms the best hard mask method (Grounded-SAM) by 16.9% mAP, demonstrating its superior ability to handle semantic ambiguity and preserve contextual relationships.

7) Comparison of Computational Costs: PuXIM avoids using computational cost processing in the inference stage for efficient retrieval. To validate its computational cost for large-scale retrieval, we compare the GFLOPs, model parameters, and average inference runtime per pair on different datasets between our method and 2 SOTA methods, *i.e.,*, SAKE [5] and ZSE-SBIR [16], and the results are listed in Table XIII. Owing to the cross-domain interaction components present in both the training and inference stages, ZSE-SBIR requires dynamically updating the feature gallery on the basis of the input sketch, resulting in a longer average runtime. The runtime increases linearly with the expansion of the image gallery. For example, since the image gallery of TU-Berlin (176081) is larger than that of Sketchy (68401), the average runtime per image increases significantly. Instead, PuXIM explores only cross-domain interactions during the training stage and enables offline retrieval, *i.e.,*, the image gallery features can be saved locally without updating them on the basis of the sketch inputs. It requires only a small

TABLE XIII
COMPARISON OF COMPUTATIONAL COST

Model	Backbone	int.		GFLOPs	Params (M)	TU-Berlin		Sketchy	
		train	infer			Runtime (ms)	mAP	Runtime (ms)	mAP
SAKE [37]	CSE-ResNet-50	✗	✗	3.90	27.6	5.12	0.475	4.65	0.547
ZSE-SBIR [34]	ViT-B/16	✓	✓	19.5	102.2	50927.95	0.542	34017.10	0.698
PuXIM (Ours)	ViT-B/16	✓	✗	17.79	87.8	5.31	0.582	4.98	0.734

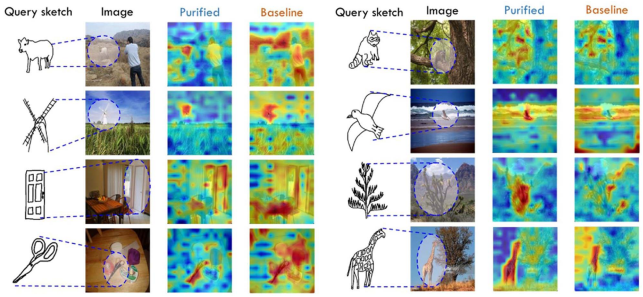

Fig. 9. Attention maps for PMM interaction. We compare the sketch–image cross-domain effects learned by the model with and without the mask during the training stage and use the sketch retrieval token [RET] as a query to display the attention maps in the inference stage. In the purified semantic space, the model can achieve more reasonable cross-domain correspondences under high semantic ambiguity conditions. In contrast, the non-purified baseline model prefers to focus on foreground regions within the image.

TABLE XIV
ABLATION STUDY ON TU-BERLIN AND SKETCHY DATASETS

Baseline	VxL	PMM		TU-Berlin Ext		Sketchy Ext	
ViT-B/16	Mask	L_{rec}	L_{int}	mAP	Prec@100	mAP	Prec@100
✓	✓			0.425	0.602	0.526	0.679
✓				0.450	0.588	0.579	0.718
✓			✓	0.468	0.584	0.599	0.717
✓		✓		0.472	0.596	0.606	0.726
✓	✓		✓	0.512	0.615	0.656	0.748
✓	✓	✓		0.530	0.657	0.680	0.788
✓	✓	✓	✓	0.582	0.673	0.734	0.810

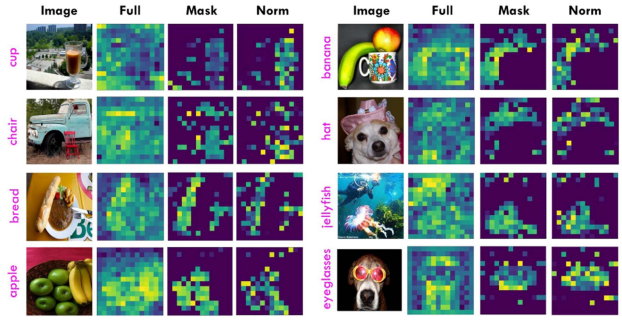

Fig. 10. Visualization of masked feature maps. We obtain attention maps by using retrieval tokens as queries to visualize feature maps. The multiple-head outputs are summed to produce the results.

computation and parameter increment, and the retrieval efficiency is comparable to that of SAKE [5]. Importantly, PuXIM's purified cross-domain interaction can significantly improve the model in zero-shot scenarios.

8) How Mask Interaction Works: Our proposed PMM interaction component is specifically designed for the training phase. However, we are still eager to ascertain whether interactions can truly learn cross-domain relationships with semantic ambiguity, thus benefiting the validation set. To investigate this, we devise the following method: using retrieval tokens of the sketch representation from the last layer as queries, we compute the similarity with pixel-level feature tokens of images to generate attention maps. Examples of sketch–image pairs with semantic ambiguity on the Sketchy Ext validation set are shown in Fig. 9. We observe that attention maps without mask operations tend to focus on the most salient regions in the images, which are easily influenced by semantically ambiguous objects such as "door" and "table", both of which exhibit highlighted areas. In contrast, attention maps with PMM interactions better align with sketch queries, effectively reducing model distraction.

9) Ablation Study: We conduct an ablation study to examine the importance of each component in PuXIM. In particular, based on Ours-full, we remove every individual component one at a time while the other parts remain, and the results are listed in Table XIV. We evaluate the hard mask matching operation, which directly zeros out irrelevant pixels without applying PMM reconstruction. While this masks ambiguous regions, it leads to lower performance than the baseline (0.425 vs. 0.450 mAP on TU-Berlin). The degradation occurs because zeroing pixels removes semantic information, disrupting important contextual cues. In zero-shot scenarios, the absence of semantic label guidance further impairs generalization, especially for semantically rich datasets. In contrast, PMM matching employs soft masking, preserving contextual information while directing the model's focus to key regions. This approach not only avoids performance degradation but also enhances generalization in zero-shot scenarios, demonstrating the importance of reconstruction in learning purified semantics. When the same no-mask features are used for cross-domain interactions as in previous methods, the model shows only marginal improvements. Instead, the proposed distraction-agnostic mask significantly improves PMM reconstruction and interaction compared with no mask. Specifically, reconstruction improves by 8% and 10.1% in mAP on TU-Berlin and Sketchy, respectively, whereas interaction improves by 6.2% and 7.7%, respectively.

10) Visualization of Normalized Masked Features: The ground truth in some of the images may have a low feature response due to redundant information, which affects the quality of the masked feature and PMM reconstruction. Hence, we compute the mean and standard deviation for the remaining tokens to obtain normalized tokens and obtain higher-quality masked features. Some examples can be seen in Fig. 10, and it can be observed that some objects obtain higher feature responses after normalization, such as "cup" and "eyeglasses". The detailed

TABLE XV
COMPARISON OF PUXIM WITH MASKED FEATURE NORMALIZATION

	norm	TU-Berlin Ext		Sketchy Ext	
		mAP	Prec@100	mAP	Prec@100
PuXIM	✗	0.560	0.653	0.708	0.782
	✓	0.582	0.673	0.734	0.810

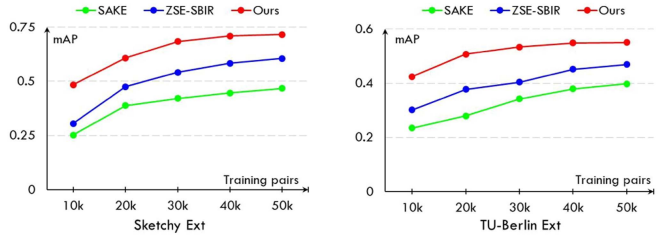

Fig. 11. Comparison of our PuXIM, SAKE, and ZSE-SBIR validation accuracies in the zero-shot setup at the training stage (the backbones are all ViT-B/16). We limit the number of training pairs to $10k$ per epoch. PuXIM is more conducive to training with reduced data requirements.

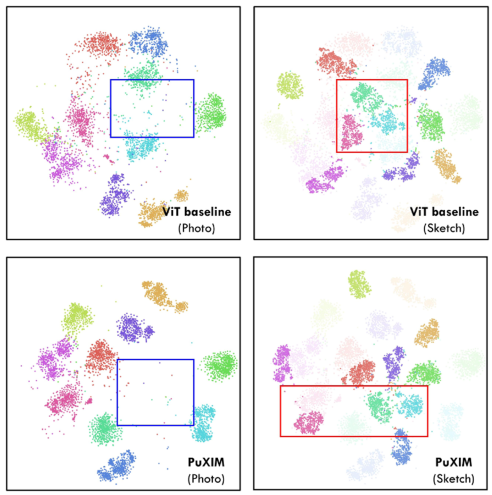

Fig. 12. Visualized t-SNE results. Image representations in the sketch semantic space are preserved and lightened for comparison. PuXIM can establish a more accurate semantic space and effectively reduce ambiguous categories and ambiguous samples (outliers).

experimental results on TU-Berlin and Sketchy are shown in Table XV, where the normalized mask yields 2.2% and 2.6% improvements in mAP, respectively.

11) Why PuXIM: The proposed framework is specifically designed for the training phase, where the focus lies on examining the advantages gained from purified sketch–image cross-domain matching. To investigate the advantages of PuXIM in zero-shot scenarios, the number of sketch–image pairs is restricted to $10k$ for each training epoch, and the validation mAP of SAKE [5] and ZSE-SBIR [16] in each epoch on the TU-Berlin and Sketchy datasets are shown in Fig. 11. For a fair comparison, both SAKE and ZS-SBIR use the same ViT-B/16 backbone. PuXIM achieves remarkable mAP with less data, whereas the performances of the other methods are substantially lower in the initial epoch with a data quantity of $10k$. We posit that this discrepancy arises from the influence of semantic ambiguity, necessitating the model to learn from a larger dataset to differentiate between semantics across different categories. In contrast, PuXIM, which is a distraction-agnostic model, can focus on matching genuine semantic correspondences between sketches and images without being influenced by ambiguous semantics. As a result, it can transfer to the zero-shot setup via reduced data.

12) Visualization of the Sketch–Image Representations: Our objective is to improve the model's ability to distinguish between sketch and image categories in semantic space. We employ t-SNE [75] to visualize the dimensionality reduction of the sketch–image representations learned by PuXIM on Sketchy Ext, which is compared with the dual encoder baseline. We randomly select results from 10 categories for display, as shown in Fig. 12. PuXIM significantly enhances the distinctiveness of outliers (semantically ambiguous and challenging samples) in semantic space, increasing the separation between different categories and correspondingly reducing misclassified samples (blue rectangular regions in the image domain and red rectangular regions in the sketch domain). Another noteworthy phenomenon is that images have more outliers than sketches do because of semantic ambiguity, which is consistent with our previous discussion. While PuXIM effectively reduces outliers (ambiguous images) through purified sketch–image matching, it also achieves better semantic space for sketches.

V. CONCLUSION AND FUTURE WORK

This paper introduces PuXIM, a novel distraction-agnostic framework for zero-shot sketch-based image retrieval (ZS-SBIR), which addresses semantic ambiguity between sketches and natural images. By integrating the visual-cross-linguistic (VxL) sampler and purified masked matching (PMM), our approach significantly improves cross-domain semantic matching. PMM, which leverages masked image representations generated by VxL, facilitates refined sketch–image pairing through enhanced reconstruction and interaction techniques. Comprehensive testing and visual analysis confirm PuXIM's superiority in ZS-SBIR, setting new benchmarks on the Sketchy and TU-Berlin datasets via basic dual encoders.

A plausible limitation of our method lies in the linguistic mask generated by CLIP, which operates at the patch level with a fixed masking rate. This approach lacks the precision needed to eliminate redundant semantics effectively, especially when the target in the image is small. Future work will explore pixel-level mask generation to increase the precision of PuXIM. Additionally, we plan to train the model on datasets with high semantic ambiguity via precise masks and transfer it to fine-grained retrieval tasks, aiming to achieve unified interclass and intraclass fine-grained retrieval with higher precision.

REFERENCES

[1] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[2] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-Stuff: Thing and stuff classes in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1209–1218.

[3] Y. Xu, X. Xu, H. Gao, and F. Xiao, "SGDM: An adaptive style-guided diffusion model for personalized text to image generation," *IEEE Trans. Multimedia*, vol. 26, pp. 9804–9813, 2024.

[4] H. Pang et al., "Heterogeneous feature alignment and fusion in cross-modal augmented space for composed image retrieval," *IEEE Trans. Multimedia*, vol. 25, pp. 6446–6457, 2023.

[5] Q. Liu, L. Xie, H. Wang, and A. L. Yuille, "Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3662–3671.

[6] T. Dutta, A. Singh, and S. Biswas, "StyleGuide: Zero-shot sketch-based image retrieval using style-guided image generation," *IEEE Trans. Multimedia*, vol. 23, pp. 2833–2842, 2021.

[7] H. Bandyopadhyay et al, "What sketch explainability really means for downstream tasks?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 10997–11008.

[8] S. Koley et al., "Picture that sketch: Photorealistic image generation from abstract sketches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6850–6861.

[9] S. Koley et al., "It's all about your sketch: Democratising sketch control in diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 7204–7214.

[10] C. Liu, S. Xu, J. Peng, K. Zhang, and D. Liu, "Towards interactive image inpainting via robust sketch refinement," *IEEE Trans. Multimedia*, vol. 26, pp. 9973–9987, 2024.

[11] Y. Shen, L. Liu, F. Shen, and L. Shao, "Zero-shot sketch-image hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3598–3607.

[12] S. K. Yelamarthi, S. K. Reddy, A. Mishra, and A. Mittal, "A zero-shot framework for sketch based image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 300–317.

[13] C. Ge et al., "Semi-transductive learning for generalized zero-shot sketch-based image retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 6, pp. 7678–7686.

[14] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–12, 2016.

[15] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, 2012.

[16] F. Lin et al., "Zero-shot everything sketch-based image retrieval, and in explainable style," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23349–23358.

[17] Z. Wang, H. Wang, J. Yan, A. Wu, and C. Deng, "Domain-smoothing network for zero-shot sketch-based image retrieval," 2021, *arXiv:2106.11841*.

[18] A. Sain et al., "Sketch3T: Test-time training for zero-shot SBIR," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7462–7471.

[19] Y. Wei et al., "Contrastive learning rivals masked image modeling in fine-tuning via feature distillation," 2022, *arXiv:2205.14141*.

[20] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5583–5594.

[21] Y. Yang et al., "Attentive mask clip," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 2771–2781.

[22] A. Sain, A. K. Bhunia, Y. Yang, T. Xiang, and Y.-Z. Song, "StyleMeUp: Towards style-agnostic sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8504–8513.

[23] A. Sain et al., "Exploiting unlabelled photos for stronger fine-grained SBIR," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6873–6883.

[24] A. Chaudhuri, A. K. Bhunia, Y.-Z. Song, and A. Dutta, "Data-free sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12084–12093.

[25] S. Koley et al., "How to handle sketch-abstraction in sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16859–16869.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25, pp. 1106–1114.

[27] X. Wang, D. Peng, P. Hu, Y. Gong, and Y. Chen, "Cross-domain alignment for zero-shot sketch-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 7024–7035, Nov. 2023.

[28] J. Tian, X. Xu, F. Shen, Y. Yang, and H. T. Shen, "TVT: Three-way vision transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 2370–2378.

[29] H. Ren et al., "ACNet: Approaching-and-centralizing network for zero-shot sketch-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 5022–5035, Sep. 2023.

[30] A. Sain et al., "Clip for all things zero-shot sketch-based image retrieval, fine-grained or not," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2765–2775.

[31] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[32] Y.-C. Chen et al., "Uniter: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 104–120.

[33] A. Miech, J.-B. Alayrac, I. Laptev, J. Sivic, and A. Zisserman, "Thinking fast and slow: Efficient text-to-visual retrieval with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9826–9836.

[34] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[36] J. Li et al., "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 9694–9705.

[37] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers," 2020, *arXiv:2004.00849*.

[38] H. Chefer, S. Gur, and L. Wolf, "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 397–406.

[39] P. Chen, Q. Li, S. Biaz, T. Bui, and A. Nguyen, "gScoreCAM: What objects is CLIP looking at," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1959–1975.

[40] Y. Li, H. Wang, Y. Duan, and X. Li, "Clip surgery for better explainability with enhancement in open-vocabulary tasks," *Pattern Recognit.*, vol. 162, 2025, Art. no. 111409.

[41] A. Radford et al., "Improving language understanding by generative pre-training," OpenAI blog, 2018.

[42] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 5998–6008.

[43] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 16433–16450.

[44] K. He et al., "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.

[45] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9653–9663.

[46] M. Chen et al., "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[47] Z. Yang et al., "Masked generative distillation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 53–69.

[48] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 611–631.

[49] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[50] S. Dey, P. Riba, A. Dutta, J. Llados, and Y.-Z. Song, "Doodle to search: Practical zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2179–2188.

[51] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2862–2871.

[52] D. Ha and D. Eck, "A neural representation of sketch drawings," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1040–1055.

[53] A. Dutta and Z. Akata, "Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5089–5098.

[54] Z. Zhang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Zero-shot sketch-based image retrieval via graph convolution network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12943–12950.

[55] X. Xu, M. Yang, Y. Yang, and H. Wang, "Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 984–990.

[56] C. Deng, X. Xu, H. Wang, M. Yang, and D. Tao, "Progressive cross-modal semantic network for zero-shot sketch-based image retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 8892–8902, 2020.
[57] U. Chaudhuri, R. Chavan, B. Banerjee, A. Dutta, and Z. Akata, "BDA-SketRet: Bi-level domain adaptation for zero-shot SBIR," *Neurocomputing*, vol. 514, pp. 245–255, 2022.
[58] O. Tursun, S. Denman, S. Sridharan, E. Goan, and C. Fookes, "An efficient framework for zero-shot sketch-based image retrieval," *Pattern Recognit.*, vol. 126, 2022, Art. no. 108528.
[59] Z. Yin, J. Yan, C. Xu, and C. Deng, "Asymmetric mutual alignment for unsupervised zero-shot sketch-based image retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 15, pp. 16504–16512.
[60] J. Yan, C. Deng, H. Huang, and W. Liu, "Causality-invariant interactive mining for cross-modal similarity learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 9, pp. 6216–6230, Sep. 2024.
[61] H. Zhang, D. Cheng, H. Jiang, J. Liu, and Q. Kou, "Task-like training paradigm in clip for zero-shot sketch-based image retrieval," *Multimedia Tools Appl.*, vol. 83, no. 19, pp. 57811–57828, 2024.
[62] E. Lyou, D. Lee, J. Kim, and J. Lee, "Modality-aware representation learning for zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 5646–5655.
[63] H. Wang, C. Deng, T. Liu, and D. Tao, "Transferable coupled network for zero-shot sketch-based image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9181–9194, Dec. 2022.
[64] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-based image retrieval by salient contour reinforcement," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1604–1615, Aug. 2016.
[65] S. Shankar et al., "Generalizing across domains via cross-gradient training," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 5099–5110.
[66] K. Pang et al., "Generalising fine-grained sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 677–686.
[67] H. Zhang et al., "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 8587–8605.
[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[69] L. Yang, K. Pang, H. Zhang, and Y.-Z. Song, "Annotation-free human sketch quality assessment," *Int. J. Comput. Vis.*, vol. 162, pp. 2743–2764, 2024.
[70] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.
[71] H. Li, X. Jiang, B. Guan, R. Wang, and N. M. Thalmann, "Multistage spatio-temporal networks for robust sketch recognition," *IEEE Trans. Image Process.*, vol. 31, pp. 2683–2694, 2022.
[72] S. Liu et al., "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 38–55.
[73] T. Ren et al., "Grounded SAM: Assembling open-world models for diverse visual tasks," 2024, *arXiv:2401.14159*.
[74] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.
[75] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

Yang Zhou received the B.Eng. degree in vehicle engineering from Anhui Agricultural University, Hefei, China. He is currently working toward the M.Eng. degree in mechanical engineering with the School of Zhejiang University, Hangzhou, China. His research focuses on computer vision.

Jingru Yang received the B.Eng. degree in machine design & manufacturing and automation from Sichuan University, Chengdu, China, and the Ph.D. degree in mechanical engineering from the School of Zhejiang University, Hangzhou, China. He is currently a Postdoc with Carnegie Mellon University, Pittsburgh, PA, USA. His research focuses on 2D and 3D computer vision.

Jin Wang received the B.Eng. degree in mechatronic engineering and the Ph.D. degree in mechanical design and theory from Zhejiang University, Zhejiang, China, in 2003 and 2008, respectively. He is currently a Professor with Zhejiang University. His research focuses on computer vision.

Kaixiang Huang received the B.Eng. degree in machine design & manufacturing and automation from Sichuan University, Chengdu, China. He is currently working toward the Ph.D. degree in mechanical engineering with the School of Zhejiang University, Hangzhou, China. His research interests include 3D computer vision and human-robot interaction.

Guodong Lu received the B.S., M.Eng., and Ph.D. degrees in applied mathematics from Zhejiang University, Zhejiang, China. He is currently a Professor with Zhejiang University. His research interests include CAD, CG, and robotics.

Shengfeng He (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Macau University of Science and Technology, Cotai, Macau, in 2009 and 2011, respectively, and the Ph.D. degree from the City University of Hong Kong, Hong Kong, in 2015. He was on the faculty of South China University of Technology, Guangzhou, China, from 2016 to 2022. He is currently an Associate Professor with the School of Computing and Information Systems, Singapore Management University, Singapore. His research interests include visual understanding and generative models. He is a senior member of CCF. He is the Lead Guest Editor of the *International Journal of Computer Vision* and an Associate Editor for *Neurocomputing*. He is the Area Chair/Senior Program Committee of ICML, AAAI, and IJCAI.