Modeling the Visual Ambiguity of Human Sketches

Yang Zhou[1] Ping Ni[1] Jin Wang[1*] Senyun Jia[1] Jingdan Yan[1] Kaixiang Huang[1]
Guodong Lu[1] Jingru Yang[2] Shengfeng He[3]
[1]Zhejiang University, Hangzhou, China
[2]Carnegie Mellon University, Pennsylvania, USA
[3]Singapore Management University, Singapore
{22260043, 22425047, dwjcom, 22360562, 22460668, kaixianghuang, lugd}@zju.edu.cn
jingruy@andrew.cmu.edu shengfenghe@smu.edu.sg

Abstract

*Human sketches provide a compact and expressive form of visual communication, but their sparse structural cues, while capturing essential object structures, introduce ambiguity because a single sketch can correspond to multiple plausible images, making cross-domain alignment uncertain and unstable. Such ambiguity fundamentally limits its sketch-based vision tasks that rely on precise sketch–image correspondence. To address this challenge, we introduce **AmbiScore**, a metric that quantifies the ambiguity of sketch-image pairs, and use Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) as a testbed to reveal how ambiguous supervision leads to performance collapse in existing methods. We further propose **DisAmb** (Disentangling Ambiguity), a framework that explicitly models and mitigates ambiguity through two components: (1) Elastic Matching, which adaptively adjusts supervision strength using AmbiScore, and (2) Purified Matching, which employs ambiguity-agnostic masks to disentangle structure and appearance via shape jigsaw and texture swapping. DisAmb establishes new benchmarks under high ambiguity and provides a robust, transferable supervisory signal for downstream sketch-guided tasks.*

1. Introduction

Learning visual representations from human sketches has long been a central focus in the computer vision community [32, 42, 56, 58]. Sketches offer a direct, controllable, and expressive interface, making them a compelling alternative to natural language [30, 31, 35, 43] or audio-based interactions [38]. Recent work has shown the potential of sketches in a variety of applications, including image retrieval [22, 25, 33, 34], semantic segmentation [17], and image generation [21, 24]. As these applications become more complex, learning robust and fine-grained correspondences between sketches and natural images becomes increasingly important.

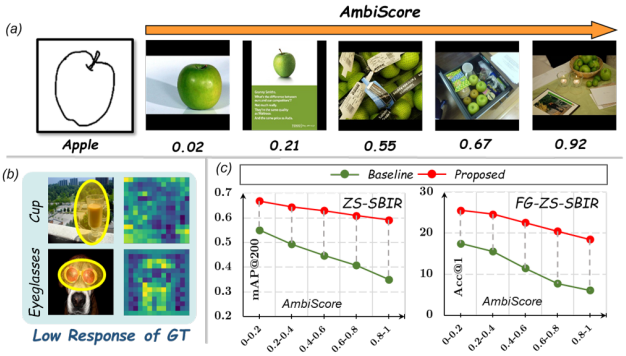

Figure 1. (a) We propose *AmbiScore*, a metric that quantifies the semantic ambiguity of sketch-image pairs. (b) Attention from the ViT **[RET]** token reveals *model distraction* on ambiguous inputs. (c) We partition the Sketchy dataset into ambiguity intervals and show that training on highly ambiguous data leads to a significant drop in ZS-SBIR and ZS-FG-SBIR performance.

Despite this promise, sketches and images differ fundamentally in visual structure. Sketches contain only sparse contours, while natural images present dense textures, colors, and backgrounds. This mismatch often introduces irrelevant cues in the image that are not reflected in the sketch. We define this mismatch as *visual ambiguity*, which raises an important question: when an image contains content unrelated to the sketch, can sketch-to-image matching remain reliable? Prior approaches in zero-shot sketch-based image retrieval (ZS-SBIR) [34, 49, 50] ignore this issue by treating all training pairs as equally informative, regardless of ambiguity.

This paper addresses the ambiguity challenge by first asking how it can be quantified and then how it affects

* Corresponding author

model behavior. Manual annotation is infeasible due to the subjective nature of ambiguity, but we observe that sketches and images often share semantic labels. This insight allows us to introduce **AmbiScore**, a metric that estimates the ambiguity level of a sketch-image pair. AmbiScore leverages vision-and-language pretraining (VLP) models [20, 55] to compute the semantic relevance between an image and its label, serving as a proxy for how well the image content matches the sketch intent, as illustrated in Fig. 1(a).

To understand how ambiguity impacts model learning, we analyze attention maps from the ViT **[RET]** token [34] during training. As shown in Fig. 1(b), we observe a *model distraction* phenomenon: attention is drawn to background elements or co-occurring objects instead of the sketch-relevant region. This behavior suggests that ambiguous training pairs can act as noisy supervision. To test this, we partition the Sketchy Extended dataset [52] into five intervals based on AmbiScore: [0.0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), and [0.8, 1.0]. As shown in Fig. 1(c), training on higher-ambiguity subsets leads to significant performance degradation, particularly in fine-grained ZS-FG-SBIR tasks, where precise matching is essential.

To overcome this, we introduce **DisAmb**, a plug-and-play training framework that explicitly accounts for visual ambiguity in sketch-image matching. DisAmb contains two key components. First, *Elastic Matching* (E-Match) uses AmbiScore to modulate supervision strength: ambiguous pairs are down-weighted, while low-ambiguity pairs are reinforced. Second, *Purified Matching* (P-Match) focuses on improving alignment for ambiguous data. Using semantic labels and open-set foundation models, we generate ambiguity-agnostic masks that retain only sketch-relevant regions. Based on these, we design two auxiliary tasks: *shape jigsaw*, which encourages structural alignment, and *texture swapping*, which focuses on semantic consistency.

We validate DisAmb across ZS-SBIR, ZS-FG-SBIR, and downstream sketch-guided tasks such as segmentation and generation. As shown in Fig. 1(c), our method consistently improves performance under various ambiguity conditions and offers a robust supervisory signal for cross-modal understanding. Overall, our contributions are threefold:

- We propose AmbiScore, a metric that quantifies sketch-image ambiguity using vision-language relevance, and reveal the impact of ambiguity on retrieval performance.
- We introduce DisAmb, a training framework with two components: E-Match, which modulates supervision using AmbiScore, and P-Match, which captures fine-grained correspondence through shape jigsaw and texture swapping.
- We demonstrate that DisAmb achieves state-of-the-art results on sketch-based retrieval benchmarks and enhances generalization to other vision tasks.

2. Related Works

2.1. Sketch-Guided Vision Tasks

Sketch-based image retrieval (SBIR) retrieves category-specific images using a query sketch [48, 51] by aligning sketch-image pairs in a shared latent space via metric learning [3, 34, 51], enabling Zero-Shot SBIR (ZS-SBIR). SAKE [37] distills ImageNet [27] semantics to improve alignment. Sketch3T [49] leverages self-supervision with contour sketches and stroke ordering for test-time training. ZSE-SBIR [34] enhances explainability via cross-attention encoders. Foundation models further boost SBIR performance to a new level [22, 25, 50]. *Sketch-based segmentation* localizes objects by a query sketch. Early methods adapt encoder-decoder architectures for segmentation [17, 46]. Recent zero-shot approaches reduce reliance on pixel annotations (mask-free): Sketch2Seg [6] trains a pixel classifier using features from a sketch-to-image diffusion model. SketchYourSeg [26] combines SBIR models and foundation models for multi-granularity mask generation. *Sketch-to-image generation* controls object morphology via sketches. GAN-based methods use multi-stage pipelines [13] or latent mapping in pre-trained GANs [21, 47]. Diffusion-based approaches often require text prompts, but ControlNet [60] enables sketch conditioning via a trainable UNet copy with zero convolutions. T2I-Adapter [40] fuses sketch features into multi-scale UNet representations. StableSketching [23] introduces an abstraction-aware framework with a sketch adapter, adaptive time-step sampling, and SBIR-guided generation.

2.2. Learning with Noisy Supervision

Learning with noisy supervision addresses the pervasive challenge of imperfect annotations in training data. Existing methods can be broadly divided into three categories. *Architecture-based methods* adjust the network architecture to model the noise transition matrix [2, 4]. *Sample-based methods* distinguish clean samples from noisy samples by treating low-loss examples as clean [1] or employing co-teaching strategies [15, 59] in which dual networks mutually select reliable samples. *Loss-based methods* enhance robustness by adjusting the loss contributions of clean and noisy samples through techniques like bootstrapping losses [44] or label correction [61], thereby mitigating the impact of mislabeled data. Recently, noisy supervision has been extended to image-text learning scenarios. NCR [11] partitions data into clean and noisy sets using network memorization effects and rectifies misaligned correspondences via adaptive co-teaching. ALBEF [30] employs momentum distillation to learn from noisy web data by generating pseudo-targets. However, prior works mainly focus on noisy supervision in vision-and-language data, while the sketch-image ambiguity problem has not been explored.

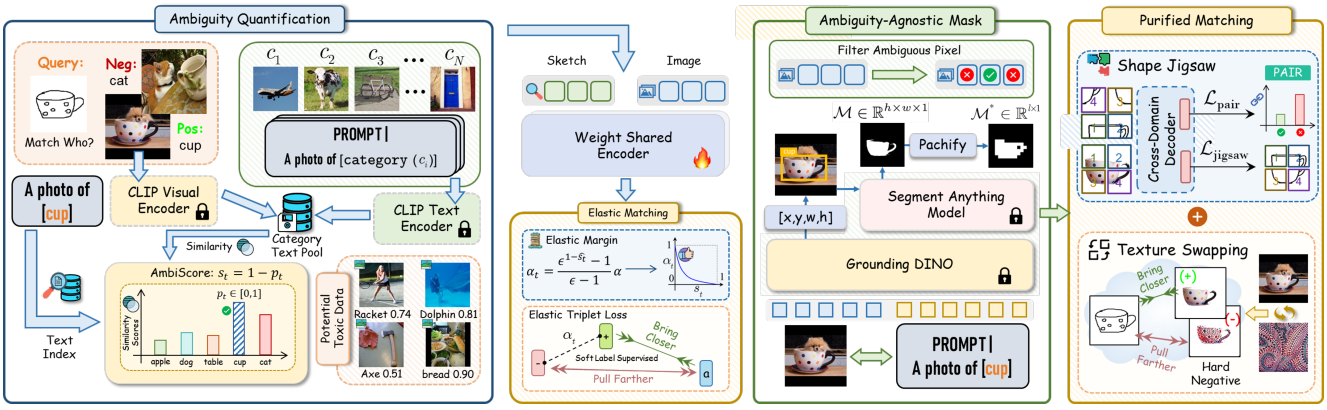

Figure 2. Our DisAmb architecture. The proposed components are applied only during training to enable efficient inference and prevent model distraction.

3. Proposed Methodology

3.1. Overview

As shown in Fig. 2, our approach builds on the SBIR framework for easy extension. Given a sketch (image) $\mathbf{x}_s$ ($\mathbf{x}_i$) and semantic label $\mathbf{T}$. We extract cross-domain embeddings via two weight-shared 12-layer ViTs (ViT-B/16) [8] as sketch (image) encoders ($\mathbf{V}_s$/$\mathbf{V}_i$). To handle semantic ambiguity, E-Match adjusts the triplet margin based on $\mathbf{T}$, while P-Match generates binarized masks $\mathcal{M}$ to filter out abundant pixels. The way of utilizing this mask is trickier, which will be described in the following sections.

3.2. Quantifying Ambiguity in Sketch-Image Pairs

We introduce an AmbiScore to evaluate the correlation between images and sketches. Given a dataset $\mathcal{D}_t$ with N categories, we manually craft a textual prompt "`a photo of [category]`" for each category, resulting in a set of category-specific prompts $\{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_N\}$. We then leverage the CLIP text encoder $\mathcal{T}(\cdot)$ to extract the corresponding text features, forming a sequence of text representations $t = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N\}$. For a given input image $\mathbf{x}_i$, we utilize the CLIP image encoder $\mathcal{I}(\cdot)$ to obtain its visual representation $\mathbf{v}_c \in \mathbb{R}^d$. To measure the semantic relevance between the image and each category prompt, we calculate the inner product between $\mathbf{v}_c$ and each text feature t_i: $p_i = \mathbf{v}_c^\top \mathbf{t}_i$ with $i = 1, 2, \ldots, N$, producing a similarity vector $\mathbf{p} = [p_1, p_2, \ldots, p_N] \in \mathbb{R}^N$. We then apply the softmax function to obtain the normalized confidence scores p_i for each category:

$$p_i = \frac{\exp(p_i/\tau)}{\sum_{j=1}^{N} \exp(p_j/\tau)}, \quad i = 1, 2, \ldots, N, \tag{1}$$

where τ is the temperature, controlling the softness of the probabilities, and $\tau = 1.5$ in our implementation. Given the ground-truth label index $y \in \{1, 2, \ldots, N\}$ corresponding to $\mathbf{x}_i$, we define the *AmbiScore* $s_t = (1 - p_t)$.

This AmbiScore quantifies the confidence in assigning the image $\mathbf{x}_i$ to its ground-truth category. Instead of manually constructing hundreds or thousands of fixed prompt combinations to calculate absolute ambiguity, AmbiScore reflects a sample's *relative* ambiguity within the dataset for better adaptation to the scale and distribution across different datasets. However, we observe that the "`person`" category is significantly more frequent in the dataset, so we additionally incorporate a prompt of "`a photo of person`" to improve generalization across diverse datasets.

3.3. Elastic Matching

We propose an elastic matching to dynamically control the supervision stringency for different ambiguous samples during the optimization process. We input $\mathbf{x}_i$ and $\mathbf{x}_s$ to the encoder $\mathbf{V}_i(\mathbf{V}_s)$ to obtain the image (sketch) features $\mathbf{v}_i(\mathbf{v}_s) \in \mathbb{R}^{l \times d}$ and its global representation $\mathbf{v}_i^{\text{cls}}(\mathbf{v}_s^{\text{cls}}) \in \mathbb{R}^d$, where d is the embedding dimension and $\mathbf{v}^{cls}$ is the class token. Prior methods adjust the global sketch-image semantics by minimizing a triplet loss with a hard margin α, which brings sketches and images of the same category closer together while distancing images of other categories. Instead, we propose a soft margin triplet loss $\mathcal{L}_{\text{EM}}$ depending on the AmbiScore that assigns larger margins to true positive pairs and smaller margins to ambiguous positive pairs. Mathematically,

$$\mathcal{L}_{\text{EM}} = \max\{d(\mathbf{v}_s^{\text{cls}}, \mathbf{v}_p^{\text{cls}}) - d(\mathbf{v}_s^{\text{cls}}, \mathbf{v}_n^{\text{cls}}) + \alpha_t, 0\}, \tag{2}$$

where $\mathbf{v}_s^{\text{cls}}$, $\mathbf{v}_p^{\text{cls}}$, and $\mathbf{v}_{\text{cls}}^n$ are the anchor sketch, positive and negative images, respectively, distance function $d(a, b) = 1 - (a \cdot b)/(||a|| \cdot ||b||)$, and α_t is the soft margin, which is adaptively determined by:

$$\alpha_t = \frac{\epsilon^{1-\hat{s}_t} - 1}{\epsilon - 1}\alpha, \tag{3}$$

where ϵ is the curve parameter and α is the margin parameter. We normalize s_t based on the gallery images by

$\hat{s}_t = (s_t - s_{min})/(s_{max} - s_{min})$, where s_{max} and s_{min} are the maximum and minimum ambiguity scores in the dataset, respectively. This operation prevents α_t from being uniformly low across highly ambiguous datasets, thereby avoiding lethargy in $\mathcal{L}_{\text{EM}}$. The objective of the above formulation is to ensure that when the s_t approaches 1, α_t will be assigned a small value, thus enabling the model to handle ambiguous pairs more flexibly and mitigating their adverse impact on the optimization process. Conversely, when the s_t approaches 0, α_t increases, enforcing a stricter separation.

3.4. Purified Matching

Ambiguity-agnostic Mask. We prefer to further leverage the ambiguous data rather than reflexively suppress them by E-Match. Therefore, we propose a hard matching strategy that more aggressively eliminates sketch-irrelevant regions in the image. We leverage the power of foundation models to select essential areas based on their semantic labels. Technically, given an input image $\mathbf{x}_i$ and its semantic label $\mathbf{T}$, Grounding DINO first localizes the object with bounding boxes, which are then used as prompts for SAM to obtain masks $\mathcal{M} \in \mathbb{R}^{h \times w \times 1}$. Its effectiveness has been validated by Ren *et al.* [45] as Grounded SAM ($\mathcal{F}(\cdot)$ in Figure 2). Based on this, we can establish the connection between semantic labels and natural images and accurately filter the sketch-irrelevant region. Considering the failure of mask extraction, we directly return the full image.

Shape Jigsaw. We downsample $\mathcal{M}$ from $\mathcal{F}$ into a patch-level binary vector $\mathbf{m} \in \{0,1\}^{\sqrt{l} \times \sqrt{l}}$ by $\mathbf{m} = \mathbb{I}\big[\text{AvgPool}_{p \times p}(\mathcal{M}) > \theta\big]$, where $\theta = 0.2$ is a threshold ratio. The mask is reshaped to $\mathcal{M}^* \in \mathbb{R}^{l \times 1}$, and the image features $\mathbf{v}_i \in \mathbb{R}^{l \times d}$ are filtered as $\mathbf{v}_{\text{mask}} = \{\mathbf{v}_i\}_{i \in S}$ with $S = \{i | \mathcal{M}^*_i = 1\}$. Now we obtain the feature sequence relevant to the sketch. We propose a shape jigsaw task based on $\mathbf{v}_{\text{mask}}$. Specifically, we leverage masked image feature $\mathbf{v}_{\text{mask}}$ and sketch feature $\mathbf{v}_s$ as the matching targets to determine whether they share the same semantic category. To enhance this process, we divide the sketch features $\mathbf{v}_s \in \mathbb{R}^{l \times d}$ into k randomly permuted segments. The shuffled sketch feature $\mathbf{v}^r_s$ and masked image feature $\mathbf{v}_{\text{mask}}$ are fed into a multi-layer transformer decoder $\mathcal{D}_r$ with cross-attention to evaluate semantic relevance:

$$\mathbf{X}c = \text{softmax}(\frac{Q(\mathbf{v}^r_s)K(\mathbf{v}_{\text{mask}})^\top}{\sqrt{d}})V(\mathbf{v}_{\text{mask}}), \quad (4)$$

where Q, K, and V are the corresponding weight matrices. We first determine whether a given pair $\Phi = (\mathbf{v}^r_s, \mathbf{v}_{\text{mask}})$ is matched by optimizing a binary classification objective:

$$\mathcal{L}_{\text{pair}} = -\big[y \log \hat{y} + (1-y)\log(1-\hat{y})\big], y \in \{0,1\}, \quad (5)$$

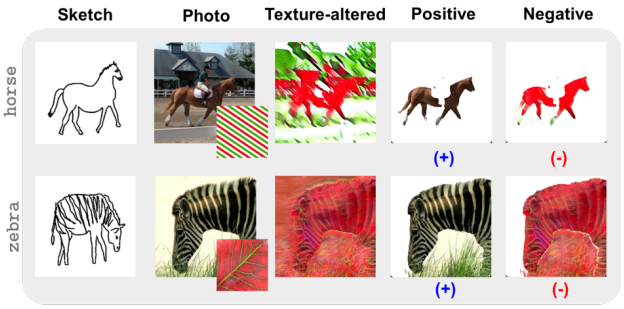

Figure 3. Texture-altered images via our masked texture swapping.

where $y \in \{0,1\}$ denotes the ground-truth label of being a matched pair and $\hat{y}$ is the predicted probability.

For positive pairs, a jigsaw loss $\mathcal{L}_{\text{jigsaw}} = \mathcal{H}(l_s, l_{gt})$ predicts the spatial order of sketch patches l_s against the ground-truth positions l_{gt}, where $\mathcal{H}(\cdot)$ denotes the multi-class cross-entropy. This auxiliary task fosters shape-aware reasoning by enforcing spatial consistency between sketches and images. Finally, the overall objective of our proposed Shape Jigsaw task is the joint optimization as $\mathcal{L}_{\text{SJ}} = \mathcal{L}_{\text{pair}} + \mathcal{L}_{\text{jigsaw}}$. Due to the high semantic diversity in natural images, establishing local sketch-image correspondences is not a trivial task, as the high complexity hinders model learning effectively. Our linguistic mask simplifies this process, making the optimization of our shape jigsaw objective feasible. We shuffle sketches rather than images, since sketches typically depict a single salient object, while images often contain multiple objects or small targets, where patch shuffling is misleading.

Texture Swapping. Images are texture-rich, and sketches are composed of stroke contour. Therefore, sketch-image matching often degenerates into a *shape-matching* task. Although intuitive, naive shape matching tends to fully overlook image texture, which is highly correlated with the object semantics [12]. We expect the model to establish semantic relationships between sketches and images, rather than learning strong shape biases. Therefore, we introduce a data augmentation trick, texture swapping.

Specifically, we construct a positive and synthetic negative image pair $(\mathbf{x}^\delta_p, \mathbf{x}^\delta_n)$. The negative sample $\mathbf{x}^\delta_n$ is derived from the positive sample $\mathbf{x}^\delta_p$ but rendered with different textures or materials. To achieve this, we apply style transfer using AdaIN [18], randomly drawing textures from the Describable Textures dataset [5] to $\mathbf{x}^\delta_p$. To ensure the model focuses exclusively on the object's intrinsic texture, we replace the input embeddings via the positive linguistic mask $\mathcal{M}^*$ by $\mathbf{v}^\delta_p = \mathcal{M}^* \odot \mathbf{v}^p_i$, $\mathbf{v}^\delta_n = \mathcal{M}^* \odot \mathbf{v}^n_i$, where $\mathbf{v}^p_i$ and $\mathbf{v}^n_i$ are the positive and negative image embeddings.

Given a triplet pair $(\mathbf{x}^s, \mathbf{x}^\delta_p, \mathbf{x}^\delta_n)$, training should decrease the distance of the sketch embedding $(\mathbf{v}_s)$ from the masked positive embedding $(\mathbf{v}^\delta_p)$, while increasing it from the synthetic negative embedding $(\mathbf{v}^\delta_n)$. Accordingly, we

devise a triplet training objective with hard margin α_{ts} as:

$$\mathcal{L}_{\text{TS}} = \max\{d(\mathbf{v}_s, \mathbf{v}_p^\delta) - d(\mathbf{v}_s, \mathbf{v}_n^\delta) + \alpha_{ts}, 0\}. \quad (6)$$

Our texture-swapping matching acts as an auxiliary hard triplet strategy, enforcing negatives to share the positive's shape while differing in texture. This enables DisAmb to focus on sketch-image semantic correspondence, handling cases with similar shapes but different semantics (e.g., "horse" and "zebra" in Fig. 3). Notably, $\mathcal{L}_{\text{TS}}$ prevents $\mathcal{L}_{\text{EM}}$ from becoming "lazy" when training on ambiguous samples, thereby facilitating effective model optimization.

With λ_{SJ} and λ_{TS} as hyperparameters. Our total training objective is

$$\mathcal{L}_{\text{DisAmb}} = \mathcal{L}_{\text{EM}} + \lambda_{\text{SJ}}\mathcal{L}_{\text{SJ}} + \lambda_{\text{TS}}\mathcal{L}_{\text{TS}}. \quad (7)$$

4. Experiments

Implementation Details. The input image and sketch size is set as 224×224 with margin parameter $\alpha = \alpha_{ts} = 1.0$ and $\epsilon = 10$. The ViT-B/16 initialized on ImageNet[12] is trained using the Adam optimizer with a learning rate $1e{-}5$ for 40 epochs on an NVIDIA A100 GPU. The segment parameter of shape jigsaw is $k = 4 \times 4$, and the transformer decoder is designed with 4 layers and 4 heads. Embedding dimensions in the latent space are 768. $\lambda_{\text{SJ}} = \lambda_{\text{TS}} = 0.5$. The AmbiScore and ambiguity-agnostic masks are calculated offline for efficient training.

Evaluation. We report mean average precision (mAP), mAP@200, and precision at the top 100 (Prec@100) and 200 (Prec@200). For FG-ZS-SBIR, following Sain *et al.* [50], accuracy is evaluated within a single category, i.e., acc.@q, denoting the proportion of top-q images matching the query sketch. We present acc.@1 and acc.@5.

Datasets. We evaluate our approach on ZS-SBIR and FG-ZS-SBIR using three standard benchmarks. (1) **Sketchy** [52] comprises 75,471 sketches across 125 categories with 100 images each. The extended version [36] adds 60,502 ImageNet [27] images. Following [57], we use 104 classes for training and 21 for zero-shot testing. (2) **TU-Berlin** [10] contains 250 categories with 80 sketches per class, extended to 204,489 images. Following [7], we split 220 classes for training and 30 for testing. (3) **QuickDraw Extended** [14] includes over 50M sketches in 345 categories. The ZS-SBIR subset provides 110 categories with 330K sketches and 204K images.

4.1. Results Analysis

Category Level ZS-SBIR. We compare the ZS-SBIR frameworks initialized on ImageNet [27] in Table 1. While existing works offer reasonable performance thanks to their respective strategies of knowledge preservation (*SAKE* [37]), test-time training (*Sketch3T* [49]), knowledge distillation (*TVT* [54]), cross-modality interaction (*ZSE-SBIR* [34]), and other setups, our DisAmb-BASE (*B-DisAmb*), built upon a concise dual-encoder architecture without any additional components during inference, still surpasses them across all benchmarks. Though DisAmb achieves only a marginal improvement over SOTA in Prec@200 on Sketchy, it surpasses prior methods by 10.7% in mAP@200, which indicates that our method produces significantly higher-quality top-200 retrieval rankings.

Figure 4. Top-5 FG-ZS-SBIR results obtained via C-DisAmb and SketchLVM on Sketchy. The green ticks denote correctly retrieved candidates, and the red crosses indicate incorrect retrievals.

We further evaluate against representative methods based on foundation models. *SketchLVM* [50], adopts a CLIP-based fine-tuning strategy, whereas *SD-PL* [22] and *SketchFusion* [25] are stable diffusion-based methods. For fair comparison, our *C-DisAmb* employs prompt fine-tuning following the implementation as *SketchLVM*, while integrating our proposed components during training. Though *SketchFusion* leverages two foundation models (CLIP and Diffusion), our *ambiguity-aware* method still achieves SOTA performance across all three datasets. Importantly, *DisAmb* acts as a training-time purification strategy, adding no inference cost and requiring no dependency on specific foundation models, making it more flexible and efficient than existing frameworks.

Cross-Category ZS-FG-SBIR. FG-ZS-SBIR is more challenging than ZS-SBIR, requiring the model to focus on fine-grained shape and texture details. For fair comparison, we evaluate models initialized on ImageNet [27], including CrossGrad [53], CC-DG [41], and ZSE-SBIR [34] in Table 2. Our method surpasses ZSE-SBIR by 4.32%, despite both approaches employing masking mechanisms. However, ZSE-SBIR primarily preserves foreground content exchange for computational efficiency, which may lead to *model distraction* in ambiguous samples. We also report results for foundation-model-based methods. While stronger backbone networks are beyond the scope of this paper, it is typically overlooked that these methods introduce *textual constraints*, which can potentially mitigate ambiguity, such as incorporating classification tasks based on text-image similarity [50]. We are the first to notice the ambiguity issue and fully leverage the potential of text prompts. Qualitative results in Fig. 4 highlight the effectiveness of DisAmb. While occasional errors occur (*e.g.*, a "tree" retrieved for a "windmill" sketch), SketchLVM frequently retrieves incorrect categories owing to its shape bias, such as

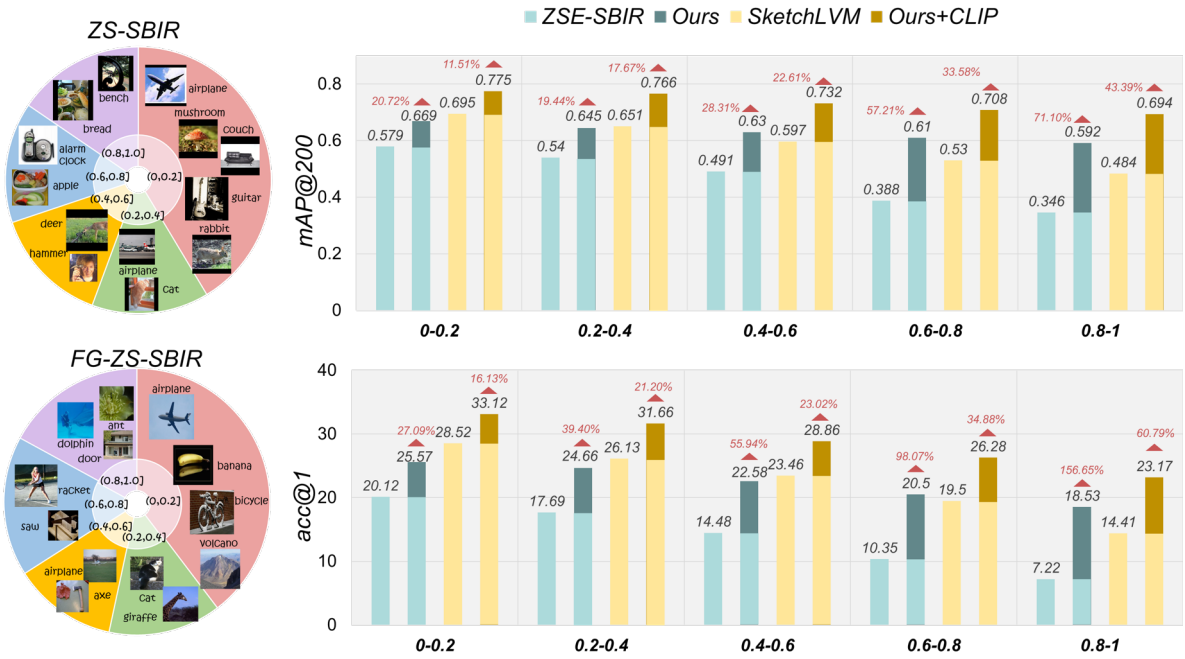

Figure 5. Comparison of our method against SOTAs across multiple subsets defined by AmbiScore intervals. For a fair comparison, we adopt the same backbones as ZSE-SBIR and SketchLVM to evaluate performance with ViT-B/16 and foundation models, respectively.

Table 1. Quantitative comparison of DisAmb against existing frameworks on ZS-SBIR datasets. "-": not reported

Methods	Sketchy		TU-Berlin		QuickDraw	
	mAP@200	Prec@200	mAP	Prec@100	mAP	Prec@200
SAKE [37]	0.497	0.598	0.475	0.599	0.130	0.179
StyleGuide [9]	0.358	0.400	0.254	0.355	-	-
Sketch3T [49]	0.575	0.624	0.507	0.648	-	-
TVT [54]	0.531	0.618	0.484	0.662	0.149	0.293
ZSE-SBIR [34]	0.525	0.624	0.542	0.657	0.145	0.216
B-DisAmb	**0.682**	**0.648**	**0.592**	0.697	**0.170**	**0.263**
SketchLVM [50]	0.723	0.725	0.651	0.732	0.202	0.388
SD-PL [22]	0.746	0.747	0.680	0.744	0.231	0.397
SketchFusion [25]	0.761	0.763	0.695	0.753	0.242	0.399
C-DisAmb	**0.812**	**0.789**	**0.707**	**0.760**	**0.245**	**0.412**

Table 2. Results on Sketchy for cross-category ZS-FG-SBIR

Methods	acc@1	acc@5	Methods	acc@1	acc@5
CrossGrad [53]	13.40	34.90	SketchLVM [50]	28.68	62.34
CC-DG [41]	22.60	49.00	SD-PL [22]	31.94	65.81
ZSE-SBIR [34]	23.97	49.52	SketchFusion [25]	33.01	67.92
B-DisAmb	**28.09**	**58.69**	***C-DisAmb***	**33.83**	**69.61**

retrieving a "bat" for a "seagull" sketch.

Ambiguous ZS-SBIR. For the ZS-SBIR and FG-ZS-SBIR datasets (Sketchy Ext and Sketchy), we analyze the ambiguity distribution in Fig. 5 by binning sketch–image pairs into five AmbiScore intervals: $s_t \in [0.0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.8]$, $(0.8, 1.0]$. Notably, only about 40% of the data in both datasets fall into the low-ambiguity range. To explicitly study the impact of ambiguity on model performance, we randomly sample 10,000 images per interval for ZS-SBIR and 1,200 images per interval for FG-ZS-SBIR. It can be observed (Fig. 5) that under high ambiguity, existing SOTAs suffer severe performance degradation, a trend particularly pronounced in FG-ZS-SBIR. Notably, Sketch-LVM exhibits limited robustness against ambiguous data due to its soft constraint based on text–image similarity, as discussed previously. However, this robustness gradually collapses when AmbiScore exceeds 0.6. In contrast, our method achieves significant gains over the baselines on both ViT-B/16 and CLIP backbones. This improvement stems from the complementary synergy of E-Match and P-Match. E-Match may become "lazy" on highly ambiguous data, while P-Match rectifies this by computing a hard triplet loss based on ambiguity-agnostic masks.

Sketch Quality. We evaluate our method in zero-shot settings across varying sketch quality levels. To this end, we introduce the Sketchy-Q dataset, comprising good (1,137 sketches), medium (1,137), and bad (917) quality samples. Following the implementation of Li *et al.* [29], we train a MobileNetV3 [16], achieving a classification accuracy of 90.85% (using 10% of sketches for validation). Dataset details are provided in the Appendix.

The trained model predicts sketch quality on three SBIR benchmarks: Sketchy, TU-Berlin, and QuickDraw (see Fig. 6). Sketchy has the highest proportion of high-quality sketches, TU-Berlin contains more medium-quality sketches, and QuickDraw features the most low-quality sketches. This setup reflects a realistic and challenging zero-shot scenario, where training and test sketch styles differ significantly. We train ZS-SBIR methods on Sketchy and TU-Berlin and evaluate them on all three datasets. Per-

formance consistently improves with sketch quality in Table 3. Our method outperforms prior work across all quality levels, exhibiting stronger adaptability to rough sketches. Due to a more accurate semantic space and fine-grained matching, it achieves a substantial 11% mAP gain over ZSE-SBIR on high-quality sketches, demonstrating robust generalization across sketch qualities.

4.2. Ablation Studies

Justifying Design Components. We evaluate our models (ViT-B/16 as the encoder), dropping one component at a time for both ZS-SBIR and FG-ZS-SBIR. As presented in Table 4, for both ZS-SBIR and FG-ZS-SBIR, using a fixed margin α_t (*w/o E-Match*) leads to a significant performance drop (0.168 mAP@200 and 10.85% acc@1 in ZS-SBIR and FG-ZS-SBIR, respectively), as the noisy supervision from ambiguous data degrades the quality of the semantic space. Removing $\mathcal{L}_{\text{pair}}$ and $\mathcal{L}_{\text{jigsaw}}$ (*w/o shape jigsaw*) and $\mathcal{L}_{\text{TS}}$ (*w/o texture swapping*) components impacts FG-ZS-SBIR more significantly, since fine-grained retrieval requires the model to capture not only overall shape but also the intricate shape, pose, and semantic correspondences between sketches and images. For FG-ZS-SBIR, texture swapping is more effective than shape jigsaw. Finally, removing the "person" prompt causes a slight performance decline (3.1% and 1.92% in ZS-SBIR and FG-ZS-SBIR, respectively), as AmbiScore measures relative ambiguity within the dataset, and the high frequency of "person" across categories weakens the model's inter-class discriminability.

For the shape jigsaw, a too-small k may not aid the model in learning shape details, while an excessively large k can make optimization difficult. We evaluate the impact of different k values on performance. From the results presented in Table 4, it can be observed that the model performs optimally when $k = 16$. However, both small and large values of k can negatively affect the model. Particularly, when k is set to 36, it can cause a performance drop.

Shape Bias in SBIR. To investigate the influence of image texture on SBIR, we extract binary masks by $\mathcal{F}$ and edge maps from the Sketchy images and train a ViT-B/16 baseline directly on these images as shown in Fig. 7. These images preserve only object outlines, with no texture. Comparing the results with the baseline trained on the original Sketchy dataset in Table 5, we observe nearly equal performance (0.466 *vs.* 0.449 *vs.* 0.479 in ZS-SBIR), indicating that SBIR frameworks exhibit a strong shape bias rather than truly focusing on the semantic content of the images. And with texture swapping enabled, the model achieved a significant 5.8% mAP@200 improvement in ZS-SBIR and 3.84% acc@1 improvement in FG-ZS-SBIR.

4.3. Task Adaptation via *Purified* SBIR Model

The SBIR model can be adapted as a supervised signal for annotation-free sketch-image matching tasks. We evaluate the adaptability of our SBIR encoder for sketch-based image segmentation (SBIS) and generation (SBIG). The objective of introducing the SBIR model is to minimize the distance between the sketch and the segmented/generated image in the latent space after encoding by the SBIR encoder. We evaluate whether the purified SBIR model can provide higher-quality supervised signals, leading to superior performance and efficient training.

Dataset. ***(i)*** For SBIS, due to the lack of available sketch-photo-mask paired datasets, we annotated 21 categories from the Sketchy dataset following prior work, adding binary mask labels to each photo. The training does not involve ground truth masks. ***(ii)*** For SBIG, we train and evaluate the model on the Sketchy dataset. We split this dataset

Figure 6. Sketchy has the highest proportion of high-quality sketches, TU-Berlin contains more medium-quality sketches, and QuickDraw features the most low-quality sketches.

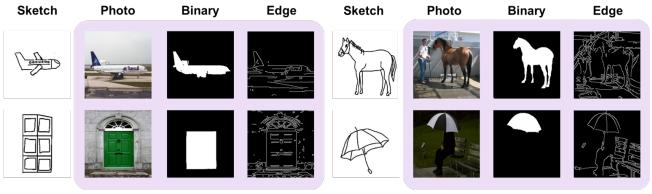

Figure 7. Images with different textures.

Table 3. mAP@200 on Sketchy and mAP on TU-Berlin for different sketch quality ZS-SBIR. The backbone of all comparative methods is ViT-B/16 for fair comparisons. "Q": QuickDraw, "T": TU-Berlin, "S": Sketchy

Methods	Sketchy			TU-Berlin		
	Q sketch	T sketch	S sketch	Q sketch	T sketch	S sketch
SAKE [37]	0.195	0.325	0.497	0.230	0.511	0.488
ZSE-SBIR [34]	0.218	0.366	0.552	0.262	0.544	0.512
DisAmb	**0.288**	**0.495**	**0.682**	**0.296**	**0.592**	**0.641**

Table 4. Ablation study and choice of k on Sketchy.

Ablation Study			Choice of k		
Methods	mAP@200	acc@1	k	mAP@200	acc@1
w/o α_t (E-Match)	0.514	17.24	4	0.603	21.74
w/o "person" prompt	0.651	26.17	9	0.642	24.55
w/o $\mathcal{L}_{\text{pair}}$ (Shape Jigsaw)	0.631	23.53	**16**	**0.682**	**28.09**
w/o $\mathcal{L}_{\text{jigsaw}}$ (Shape Jigsaw)	0.616	22.01	25	0.661	25.35
w/o $\mathcal{L}_{\text{TS}}$ (Texture Swapping)	0.635	21.28	36	0.580	20.13
Ours-full	**0.682**	**28.09**			

Table 5. Model performance on different texture images.

Type	ZS-SBIR		FG-ZS-SBIR	
	mAP@200	Prec@200	acc@1	acc@5
Mask	0.466	0.410	14.08	36.11
Edge map	0.449	0.398	13.79	35.27
Image	0.479	0.423	14.74	37.92
Image w/ $\mathcal{L}_{\text{TS}}$	**0.537**	**0.522**	**18.58**	**44.21**

Figure 8. Segmentation map generated by SkechYourSeg and our method (left). Image generated by StableSketch and our method (right).

into a 9:1 ratio for training and evaluation.

Training & Evaluation. We train segmentation and generation models following the implementations of *SketchYourSeg* [26] and *StableSketch* [24], with the key difference that we incorporate our pre-trained SBIR model as a supervisory signal. The segmentation task utilizes a category-level SBIR model. We use *mIoU* and *pAcc* as evaluation metrics. The generation task employs a fine-grained SBIR model. We evaluate the generation quality and sketch-fidelity with two metrics: *Frechet Inception Distance-InceptionV3 (FID-I)* [19] and *CLIP (FID-C)* [28] following prior works [39, 40, 60]. Lower values of FID-I and FID-C depict better generation quality. For fair comparison, the SBIR model applies ViT-B/16 in all experiments.

Sketch-Based Image Segmentation. Based on the quantitative and qualitative results presented, we find that the model using a standard SBIR loss for segmentation supervision tends to focus only on salient regions of the image. The quantitative results in Table 6 show that the segmentation pAcc using the purified SBIR model is significantly higher than the baseline by 11.57%. Fig. 8 shows several failure cases produced by prior work: the target "umbrella" is misdirected as "person". In contrast, our purified SBIR model provides a more reliable and fine-grained supervisory signal, yielding higher-quality segmentation masks.

Sketch-Based Image Generation. In sketch-based image generation, although StableSketch can produce reasonable color schemes and styles, its outputs still suffer from semantic errors and structural distortions. For example, clear semantic misinterpretations are observed in the generated "butterfly" and "dolphin" in Fig. 8. In contrast, thanks to the ambiguity-agnostic nature of our SBIR model's supervisory signal, the images generated are closer to real-world scenes and exhibit fewer deformations. By leveraging shape and texture-level awareness techniques, our method accurately captures the intended semantics without compromising object structure. The quantitative results in Table 6 demonstrate that incorporating a more precise SBIR supervisory signal outperforms existing methods on both FID-I (26.21 *vs.* 21.05) and FID-C (12.82 *vs.* 8.63).

Table 6. Quantitative comparison on the Sketchy for segmentation and generation.

Segmentation			Generation		
Methods	mIOU↑	pAcc.↑	**Methods**	FID-I↓	FID-C↓
SketchYourSeg [26]	65.68	68.29	StableSketch [24]	26.21	12.82
Ours	**71.27**	**79.86**	**Ours**	**21.05**	**8.63**

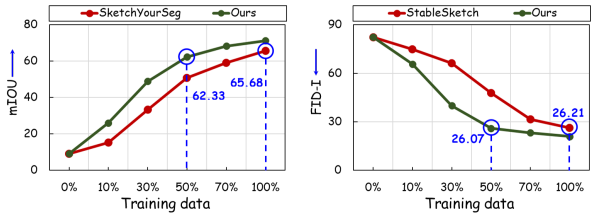

Figure 9. Comparison of our SBIR model with SOTAs under different training dataset volumes.

Data Volume. We further evaluate the effectiveness of our method under different data volumes in Fig. 9. As the volume of training data increases, the advantage of our approach becomes more significant. Notably, in both sketch-based image segmentation and generation, our method achieves an mIOU of 62.33 and an FID-I of 26.07, respectively, using 50% of the training data. These results demonstrate that the purified SBIR supervision signal is highly data efficient, enabling performance that matches or even exceeds that of SOTAs while requiring nearly half the amount of training data.

5. Conclusion

This paper revisits sketch–image matching through the lens of ambiguity, proposing effective strategies to quantify and mitigate it. We introduce elastic and purified matching to address noisy supervision, achieving consistent gains over state-of-the-art methods across multiple benchmarks. Moreover, by extending our purified SBIR encoder to sketch-based segmentation and generation, our framework attains better performance and enables more efficient adaptation to downstream sketch–guided vision tasks. A limitation is that hand-crafted prompts may cause minor fluctuations in AmbiScore, leading to inconsistent ambiguity judgments. Future work will explore human feedback or multimodal large language models (MLLM) for more reliable ambiguity evaluation.

Acknowledgment

This work is supported by the National Key R&D Program of China (No.2022YFB3303102), Robotics Institute of Zhejiang University under Grant K11808 and K11811.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR, 2019. 2

[2] Alan Joseph Bekker and Jacob Goldberger. Training deep neural-networks based on unreliable labels. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2682–2686. IEEE, 2016. 2

[3] Abhra Chaudhuri, Ayan Kumar Bhunia, Yi-Zhe Song, and Anjan Dutta. Data-free sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12084–12093, 2023. 2

[4] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1431–1439, 2015. 2

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 4

[6] Xin Dai, Haoge Deng, Ke Li, and Yonggang Qi. Sketch2seg: Sketch-based image segmentation with pre-trained diffusion model. In *International Conference on Pattern Recognition*, pages 36–50. Springer, 2024. 2

[7] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2179–2188, 2019. 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[9] Titir Dutta and Soma Biswas. Style-guided zero-shot sketch-based image retrieval. In *BMVC*, page 9, 2019. 6

[10] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Transactions on graphics (TOG)*, 31 (4):1–10, 2012. 5

[11] Zerun Feng, Zhimin Zeng, Caili Guo, Zheng Li, and Lin Hu. Learning from noisy correspondence with tri-partition for cross-modal matching. *IEEE Transactions on Multimedia*, 26:3884–3896, 2023. 2

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018. 4, 5

[13] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1171–1180, 2019. 2

[14] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017. 5

[15] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018. 2

[16] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 6

[17] Conghui Hu, Da Li, Yongxin Yang, Timothy M Hospedales, and Yi-Zhe Song. Sketch-a-segmenter: Sketch-based photo segmenter generation. *IEEE transactions on image processing*, 29:9470–9481, 2020. 1, 2

[18] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 4

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 8

[20] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[21] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Picture that sketch: Photorealistic image generation from abstract sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6850–6861, 2023. 1, 2

[22] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Text-to-image diffusion models are great sketch-photo matchmakers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16826–16837, 2024. 1, 2, 5, 6

[23] Subhadeep Koley, Ayan Kumar Bhunia, Deeptanshu Sekhri, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. It's all about your sketch: Democratising sketch control in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7204–7214, 2024. 2

[24] Subhadeep Koley, Ayan Kumar Bhunia, Deeptanshu Sekhri, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. It's all about your sketch: Democratising sketch control in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7204–7214, 2024. 1, 8

[25] Subhadeep Koley, Tapas Kumar Dutta, Aneeshan Sain, Pinaki Nath Chowdhury, Ayan Kumar Bhunia, and Yi-Zhe Song. Sketchfusion: Learning universal sketch features through fusing foundation models. *arXiv preprint arXiv:2503.14129*, 2025. 1, 2, 5, 6

[26] Subhadeep Koley, Viswanatha Reddy Gajjala, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, Ayan Kumar Bhunia, and Yi-Zhe Song. Sketchyourseg: Mask-free subjective image segmentation via freehand sketches. *arXiv preprint arXiv:2501.16022*, 2025. 2, 8

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2, 5

[28] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022. 8

[29] Hanhui Li, Xudong Jiang, Boliang Guan, Ruomei Wang, and Nadia Magnenat Thalmann. Multistage spatio-temporal networks for robust sketch recognition. *IEEE Transactions on Image Processing*, 31:2683–2694, 2022. 6

[30] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 2

[31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[32] Lei Li, Changqing Zou, Youyi Zheng, Qingkun Su, Hongbo Fu, and Chiew-Lan Tai. Sketch-r2cnn: an rnn-rasterization-cnn architecture for vector sketch recognition. *IEEE transactions on visualization and computer graphics*, 27(9):3745–3754, 2020. 1

[33] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 1

[34] Fengyin Lin, Mingkang Li, Da Li, Timothy Hospedales, Yi-Zhe Song, and Yonggang 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*, pages 23349–23358, 2023. 1, 2, 5, 6, 7

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[36] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2871, 2017. 5

[37] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3662–3671, 2019. 2, 5, 6, 7

[38] Shentong Mo and Pedro Morgado. Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27186–27196, 2024. 1

[39] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 8

[40] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. 2, 8

[41] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Generalising fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 677–686, 2019. 5, 6

[42] Zhiyu Qu, Yulia Gryaditskaya, Ke Li, Kaiyue Pang, Tao Xiang, and Yi-Zhe Song. Sketchxai: A first look at explainability for human sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23327–23337, 2023. 1

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[44] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014. 2

[45] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4

[46] Pau Riba, Sounak Dey, Ali Furkan Biten, and Josep Llados. Localizing infinity-shaped fishes: Sketch-guided object localization in the wild. *arXiv preprint arXiv:2109.11874*, 2021. 2

[47] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 2

[48] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8504–8513, 2021. 2

[49] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch3t: Test-time training for zero-shot sbir. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7462–7471, 2022. 1, 2, 5, 6

[50] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe 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*, pages 2765–2775, 2023. 1, 2, 5, 6

[51] Aneeshan Sain, Ayan Kumar Bhunia, Subhadeep Koley, Pinaki Nath Chowdhury, Soumitri Chattopadhyay, Tao Xiang, and Yi-Zhe Song. Exploiting unlabelled photos for stronger fine-grained sbir. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6873–6883, 2023. 2

[52] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 2, 5

[53] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018. 5, 6

[54] Jialin Tian, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. Tvt: Three-way vision transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2370–2378, 2022. 5, 6

[55] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 2

[56] Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, Zhanyu Ma, and Jun Guo. Sketchmate: Deep hashing for million-scale human sketch retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8090–8098, 2018. 1

[57] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 300–317, 2018. 5

[58] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016. 1

[59] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International conference on machine learning*, pages 7164–7173. PMLR, 2019. 2

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2, 8

[61] Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, and Chao Chen. Error-bounded correction of noisy labels. In *International Conference on Machine Learning*, pages 11447–11457. PMLR, 2020. 2