# PRESERVING MANIFOLD CAPACITY: A GEOMETRY-AWARE DUAL-PROMPT FRAMEWORK FOR OPEN-WORLD CROSS-MODAL RETRIEVAL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

In large-scale cross-modal data mining, constructing an embedding space that generalizes to open-set distributions is a fundamental challenge. We investigate a geometric failure mode, termed "Feature Space Cannibalization," in which excessive adaptation to seen categories may reduce the representational flexibility required for generalization to unseen categories. This paper proposes a Geometry-Aware Dual-Prompting (GADP) framework to address this topology distortion. We introduce Modality-Adaptive Contextual Tokens as distribution anchors to adapt heterogeneous inputs (*e.g.*, sparse sketches and dense images) to a pre-trained manifold without catastrophic forgetting. Crucially, we reinterpret an established angular-margin objective as a compactness regularizer for zero-shot cross-modal retrieval. By synergizing global alignment with angular margin regularization, we actively compact the distributions of seen classes, encouraging compact seen-class distributions and preserving representational flexibility for unseen categories. Experiments on three datasets under four standard evaluation settings demonstrate the effectiveness of GADP. In particular, the method reports 34.5% mAP@all on QuickDraw Ext. under our evaluation setup.

## 1 INTRODUCTION

Cross-Modal Information Retrieval (CMIR) lies at the core of modern knowledge discovery and data mining systems. User intent is often conveyed via sparse abstract inputs, while targets form large-scale, dense, highly heterogeneous spaces. Practical scenarios including e-commerce sketch-based visual search and metaverse content creation require retrieving fine-grained concrete objects from incomplete structural cues. We adopt Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) (Lyou et al., 2024) as a representative testbed for sparse-to-dense cross-modal matching (Yang et al., 2024), where structural abstract queries and textured dense targets have disjoint training and test label spaces.

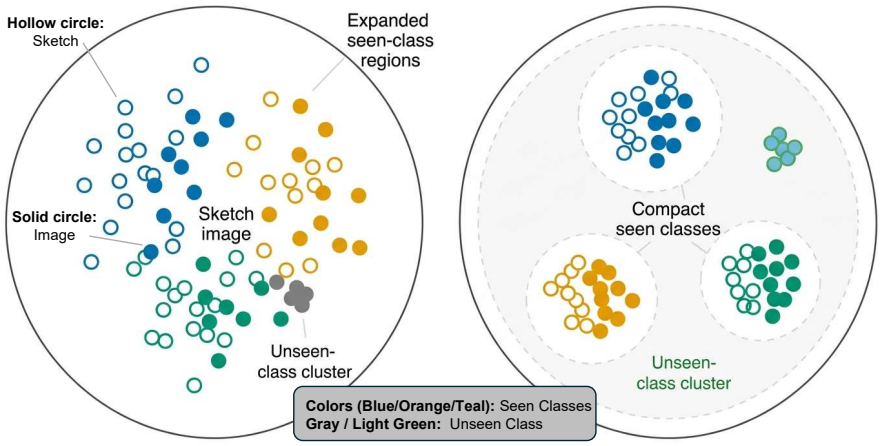
ZS-SBIR serves as an ideal platform to investigate open-world embedding generalization. Large vision-language models such as CLIP (Radford et al., 2021) provide a shared semantic manifold, yet direct deployment yields a "Generalization Paradox" (Wortsman et al., 2022). Aggressive cross-modal alignment on seen categories distorts pre-trained topological structures, resulting in overfitting and degraded discriminability for unseen classes.

Existing approaches overlook Manifold Capacity. Within the finite hyperspherical embedding space used by contrastive models (*e.g.*, CLIP's unit hypersphere (Ge et al., 2023)), categories compete for limited geometric volume. We identify the phenomenon "Feature Space Cannibalization" (Lin et al., 2020). Traditional fine-tuning or unconstrained prompt learning drives seen classes to greedily expand decision boundaries, consuming vacant regions reserved for unseen concepts. Once latent space is saturated, novel representations overlap with known ones and break open-world retrieval. As visualized in Fig. 1, conventional adaptation produces crowded embeddings, whereas GADP maintains representational flexibility by compacting seen-class distributions.

To tackle this issue, we present Geometry-Aware Dual-Prompting (GADP), casting ZS-SBIR as a manifold topology optimization problem under the "reserve-by-compacting" principle. Full fine-tuning harms topological integrity; our GADP acts as lightweight surgical optimization: prompts



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




(a) Conventional adaptation (b) GADP: reserve by compacting

Figure 1: Illustration of Feature Space Cannibalization. (a) Conventional adaptation expands the occupied regions of seen classes, leaving little representational flexibility for unseen categories. (b) Our Geometry-Aware Dual-Prompting (GADP) follows a reserve-by-compacting principle that encourages compact seen-class distributions, thereby preserving latent manifold capacity for unseen classes.

bridge modality gaps, and geometric capacity regularization enforces compactness. We posit a valid zero-shot embedding space requires three geometric properties: (1) Global Isometry: sketches and images are aligned without corrupting CLIP's pre-trained topology; (2) Local Rectification: ambiguous hard negatives are pushed away from decision boundaries; (3) Capacity Reservation: minimizing seen-class intra-class variance to reserve latent space for unseen categories.

GADP implements a unified geometric optimization pipeline. First, Modality-Adaptive Contextual Tokens work as lightweight distribution anchors to map heterogeneous inputs toward CLIP's active manifold center and decouple structural abstraction from semantics. Second, we adopt dual-constraint optimization: global contrastive loss guarantees cross-modal alignment, and Hypersphere Energy Regularization with angular margins suppresses angular dispersion of seen classes. Such compactness reserves representation flexibility for unknown concepts, without theoretically guaranteeing fixed vacant regions for new categories. Our main contributions are summarized below:

- We identify and empirically analyze "Feature Space Cannibalization" in ZS-SBIR, demonstrating that over-adaptation on seen data hurts generalization to unseen categories.
- We propose GADP, integrating class-agnostic modality-specific prompts with cross-modal alignment, hard-negative separation and angular compactness regularization.
- Experiments on three datasets with four evaluation protocols achieve competitive retrieval performance and verify the correlation between seen-class compactness and unseen-class generalization.

## 2 Related Work

### 2.1 High-Dimensional Embedding & Manifold Learning

Embedding learning is central to modern data mining. While metric learning approaches like Triplet Loss focus on local pairwise distances, they often neglect the global volumetric constraints of the embedding space. Recent works in manifold learning suggest that the "Compactness-Separability" trade-off is governed by the geometry of the hypersphere. Our work bridges the gap between these theoretical insights and practical cross-modal retrieval, explicitly modeling the "Manifold Capacity" to ensure open-set generalization.

### 2.2 ZS-SBIR: From Generative Synthesis to Manifold Alignment

Early Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) methods primarily treated the modality gap as a synthesis problem (Sohn et al., 2015). Generative approaches, such as CycleGAN-based translation (Zhu et al., 2020), attempted to bridge the pixel-level discrepancy by hallucinating textures onto sketches. However, from a geometric perspective, these methods introduce synthetic noise that distorts the intrinsic structural topology, leading to unstable retrieval boundaries.

Subsequent discriminative methods shifted focus to Global Isometry—aligning sketch and image embeddings in a shared semantic space via metric learning or adversarial training (Dutta et al., 2020). While effective for seen categories, these approaches often suffer from what we define as "Feature Space Cannibalization" (Kodirov et al., 2017): excessive alignment on training classes expands their intra-class variance, inadvertently "eating up" the latent space required for unseen categories. Unlike these methods, our work pivots from simple alignment to Manifold Topology Optimization, aiming to rectify these distortions to preserve the open-world geometry.

### 2.3 Prompt Learning as Topological Anchoring

With the advent of foundational Vision-Language Models like CLIP, Prompt Learning has emerged as a parameter-efficient alternative to fine-tuning. Methods like CoOp (Zhou et al., 2022a), CoCoOp (Zhou et al., 2022b), and MaPLe (Khattak et al., 2023) introduce learnable tokens to adapt pre-trained manifolds to downstream tasks. However, existing prompt strategies are predominantly designed for single-modality domain adaptation (*e.g.*, natural image classification), addressing style shifts rather than the fundamental structural-semantic abstraction gap in ZS-SBIR.

Blindly applying these prompts can lead to topological collapse, where the delicate open-world knowledge of CLIP is overwritten by domain-specific features. In our framework, we reinterpret prompts not merely as adapters, but as "Distribution Anchors". By employing modality-adaptive tokens, we decouple structural abstraction (sketches) from semantic context (images), anchoring the input distributions to the stable regions of the CLIP manifold without disrupting its pre-trained generalization capability.

### 2.4 Open-Set Regularization and Capacity Reservation

The most critical challenge in ZSL (Yang et al., 2024) is maintaining discriminability for unseen classes. Traditional objectives, such as Softmax and standard Contrastive Loss, focus on maximizing the separation of seen classes but neglect the volumetric distribution of the embedding space (Wang & Isola, 2020).

Recent advances in Angular Margin losses, such as CosFace (Liu et al., 2021a) and ArcFace/AM-Softmax (Wang et al., 2018), have introduced geometric constraints to enhance boundary sharpness. While originally developed for closed-set face recognition, we identify a profound connection between these margins and Open-Set Capacity. Recent theoretical works suggest that compacting intra-class variance on a hypersphere implicitly maximizes the "void space" available for novel categories.

Building on this insight, we formally adopt AM-Softmax as a Hypersphere Energy Regularizer for ZS-SBIR. Unlike prior works that use it solely for discrimination, we explicitly leverage it to minimize the "angular footprint" of seen classes, thereby reserving the essential Manifold Capacity required for robust zero-shot generalization.

## 3 Preliminaries

We formulate Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) as a Manifold Topology Optimization problem under open-set constraints. Let $\mathcal{S} = \{s_i\}_{i=1}^{N_s}$ and $\mathcal{I} = \{x_j\}_{j=1}^{N_i}$ denote the sketch and image domains, respectively. The training set consists of seen categories $\mathcal{C}_{seen}$, while the testing set comprises disjoint unseen categories $\mathcal{C}_{unseen}$.

The goal is to learn a parametric mapping function $F_\theta : \mathcal{S} \cup \mathcal{I} \to \mathcal{M}$ that projects heterogeneous inputs onto a shared hyperspherical manifold $\mathcal{M} \subset \mathbb{R}^d$ (where $||\mathbf{z}||_2 = 1$). This setting is consistent

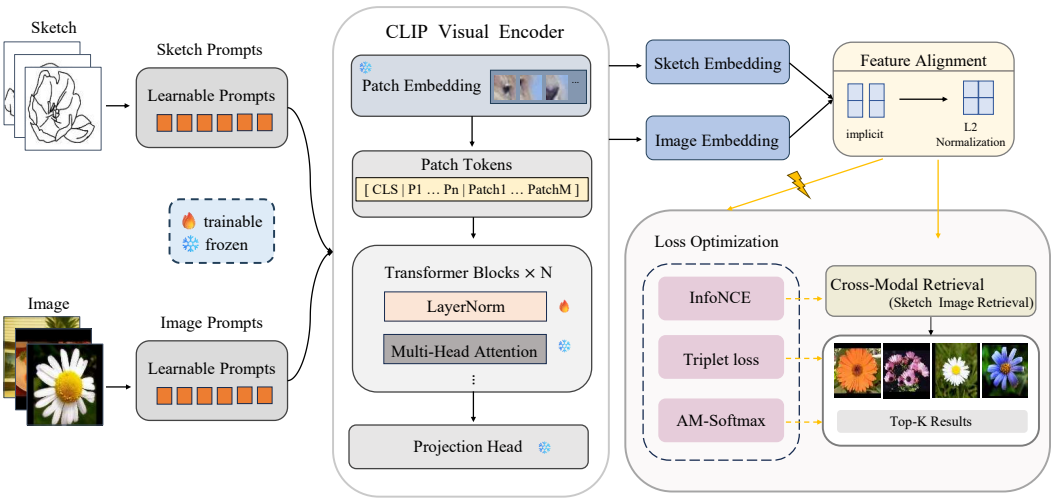



Figure 2: Overall framework. Modality-adaptive prompts are concatenated with sketch and image inputs and passed through a shared CLIP visual encoder. The backbone remains frozen except for prompt tokens and LayerNorm parameters. The resulting $\ell_2$-normalized embeddings are optimized using Global Isometry ($\mathcal{L}_{align}$), Local Rectification ($\mathcal{L}_{rect}$), and Capacity Reservation ($\mathcal{L}_{center}$ and $\mathcal{L}_{cap}$). At inference, gallery images are ranked by cosine similarity to return the top-$K$ results.

with modern contrastive foundation models (*e.g.*, CLIP), where embeddings are typically normalized and lie on the unit hypersphere.

To make the notion of "Manifold Capacity" explicit, we provide the following geometric definitions. Let the embedding space be the unit hypersphere $\mathbb{S}^{d-1}$. Its total surface area is given by:

$$A_{total} = \frac{2\pi^{d/2}}{\Gamma(d/2)}. \tag{1}$$

We interpret $A_{total}$ as the finite geometric capacity of the embedding space. For a seen class $c \in \mathcal{C}_{seen}$, its occupied region on $\mathbb{S}^{d-1}$ is defined as:

$$\Omega_c = \{z \in \mathbb{S}^{d-1} | P(c|z) \geq \tau\}, \tag{2}$$

where $\tau$ is a confidence threshold, $P(c|z)$ denotes the posterior induced by the angular-margin classifier. Geometrically, $\Omega_c$ can be approximated as a hyperspherical cap characterized by an angular radius $\theta_c$ around the class prototype. We denote its surface measure as $Vol(\Omega_c)$, which captures the class's angular footprint. The latent capacity available for unseen categories, referred to as the void volume, is defined as:

$$V_{void} = A_{total} - \sum_{c \in \mathcal{C}_{seen}} Vol(\Omega_c). \tag{3}$$

A robust open-world embedding requires $V_{void}$ ¿ 0, so that novel categories can be mapped into non-overlapping regions.

We define the *"Feature Space Cannibalization"* risk as the uncontrolled expansion of the intra-class angular variance $\sigma_\theta^2$ of seen classes, which increases $\theta_c$ and thus enlarges $Vol(\Omega_c)$. In the extreme case, optimizing only for retrieval accuracy on $C_{seen}$ can lead to $\sum_{c \in \mathcal{C}_{seen}} Vol(\Omega_c) \to A_{total}$, driving $V_{void} \to 0$, driving $V_{void} \to 0$. Under this condition, any mapping of an unseen category must inevitably overlap with occupied regions, resulting in retrieval collisions and degraded zero-shot generalization.

Based on the above definitions, our objective can be formulated as preserving manifold capacity by explicitly bounding the total occupied volume of seen classes, while maintaining sufficient cross-modal alignment:

$$\min_{\theta} \sum_{c \in \mathcal{C}_{seen}} \text{Vol}(\Omega_c) \quad \text{s.t.} \quad \mathcal{D}_{align}(\mathcal{S}, \mathcal{I}) < \epsilon, \tag{4}$$

where $\mathcal{D}_{align}$ denotes the cross-modal alignment error. This objective highlights the core principle of our framework: to maximize the reserved void volume for the open world, the angular footprint of seen classes must be strictly compacted rather than allowed to sprawl across the manifold. In practice, $\text{Vol}(\Omega_c)$ serves as a conceptual geometric proxy rather than an explicitly computed quantity; our optimization minimizes it implicitly via angular margin constraints.

# 4 Proposed Method

Figure 2 presents an overview of the proposed Geometry-Aware Dual-Prompting (GADP) framework. Modality-adaptive prompts are injected into a shared, largely frozen CLIP visual encoder to produce normalized sketch and image embeddings. Training is formulated as a joint geometric optimization with three objectives: Global Isometry for cross-modal alignment, Local Rectification for boundary refinement, and Capacity Reservation for compacting seen-class distributions on the hypersphere. Together, these objectives preserve geometric capacity for unseen classes and mitigate feature space cannibalization during cross-modal retrieval. In the following sections, we introduce each component of GADP in detail.

## 4.1 Distribution Anchoring via Dual-Prompting

Directly fine-tuning the visual encoder $f_\psi$ (*e.g.*, CLIP ViT-L/14) risks distorting the pre-trained topology. We introduce Modality-Adaptive Contextual Tokens to act as distribution anchors.

We model the input sketch $s$ and image $x$ as samples from divergent distributions $P_\mathcal{S}$ and $P_\mathcal{I}$. We introduce learnable prompt vectors $\mathbf{v}_s, \mathbf{v}_i \in \mathbb{R}^{K\times D}$ to adapt these distributions to the CLIP manifold center. The encoding process is formulated as a conditional probability adaptation:

$$\mathbf{z}_s = f_\psi([s \oplus \mathbf{v}_s]), \quad \mathbf{z}_i = f_\psi([x \oplus \mathbf{v}_i]), \tag{5}$$

where $[\cdot \oplus \cdot]$ denotes token concatenation. Crucially, we freeze the majority of the backbone parameters $\psi$ and only update the LayerNorm parameters $\psi_{LN}$ and prompts $\mathbf{v}$. From a Bayesian perspective, this mechanism acts as a *structure-semantic decoupling* filter:

$$P(\mathbf{z}|s, \mathbf{v}_s) \approx P(\mathbf{z}|s) \cdot \frac{P(\mathbf{v}_s|\text{structure})}{P(\mathbf{v}_s)}. \tag{6}$$

This anchors the modality-specific noise while preserving the global semantic topology necessary for generalization.

## 4.2 Global Isometry via Information Maximization

To satisfy the global alignment constraint ($\mathcal{D}_{align} < \epsilon$), we maximize the mutual information between the sketch and image embeddings of the same semantic class. Following the InfoNCE paradigm, we approximate the lower bound of mutual information $I(\mathbf{z}_s; \mathbf{z}_i)$.

Let $\mathbf{z}_s^i$ and $\mathbf{z}_i^j$ be the $L_2$-normalized embeddings. The cosine similarity matrix is defined as $M_{ij} = (\mathbf{z}_s^i)^\top \mathbf{z}_i^j / \tau$. We define the joint probability distribution of matched pairs as:

$$p(i, j) = \frac{\exp((\mathbf{z}_s^i)^\top \mathbf{z}_i^j / \tau)}{\sum_{k=1}^N \exp((\mathbf{z}_s^i)^\top \mathbf{z}_i^k / \tau)}. \tag{7}$$

The global alignment loss $\mathcal{L}_{align}$ is defined as the symmetric cross-entropy loss to maximize the log-likelihood of matched pairs:

$$\mathcal{L}_{s\to i} = -\frac{1}{N}\sum_{i=1}^N \log \frac{\exp((\mathbf{z}_s^i)^\top \mathbf{z}_i^+ / \tau)}{\sum_{k=1}^N \exp((\mathbf{z}_s^i)^\top \mathbf{z}_i^k / \tau)}, \tag{8}$$

$$\mathcal{L}_{align} = \frac{1}{2}(\mathcal{L}_{s\to i} + \mathcal{L}_{i\to s}). \tag{9}$$

Minimizing Eq. (9) is theoretically equivalent to maximizing the mutual information lower bound:

$$I(\mathbf{z}_s; \mathbf{z}_i) \geq \log(N) - \mathcal{L}_{align}. \tag{10}$$

This ensures Global Isometry, aligning the semantic manifolds of sketches and images globally.

### 4.3 LOCAL RECTIFICATION VIA HARD-NEGATIVE SEPARATION

Global alignment is insufficient to resolve "Boundary Ambiguity" in high-dimensional space. Hard negatives often form folded manifolds where the geodesic distance is distorted. We employ a Triplet Constraint to rectify the local curvature.

For an anchor sketch $\mathbf{z}_s$, a positive image $\mathbf{z}_i^+$, and a hard-negative image $\mathbf{z}_i^-$, we enforce a margin $\delta$ in the geodesic space. Since we operate on a unit hypersphere, the Euclidean distance constraint implies:

$$||\mathbf{z}_s - \mathbf{z}_i^+||_2^2 + \delta < ||\mathbf{z}_s - \mathbf{z}_i^-||_2^2 \tag{11}$$

Substituting $||\mathbf{a} - \mathbf{b}||_2^2 = 2(1 - \cos\langle \mathbf{a}, \mathbf{b} \rangle)$, we derive the angular constraint:

$$\cos\langle \mathbf{z}_s, \mathbf{z}_i^- \rangle < \cos\langle \mathbf{z}_s, \mathbf{z}_i^+ \rangle - \frac{\delta}{2} \tag{12}$$

Equivalently, the positive-pair similarity must exceed the negative-pair similarity by at least $\delta/2$. The Local Rectification Loss is formulated as:

$$\mathcal{L}_{rect} = \frac{1}{N} \sum_{n=1}^{N} \max\left(0, \cos\langle \mathbf{z}_s^n, \mathbf{z}_i^{-n} \rangle - \cos\langle \mathbf{z}_s^n, \mathbf{z}_i^{+n} \rangle + m_{loc}\right) \tag{13}$$

where $m_{loc} = \delta/2$. This objective enforces an explicit similarity margin between positive and hard-negative pairs, thereby reducing local retrieval ambiguity.

### 4.4 HYPERSPHERE CAPACITY RESERVATION

This section addresses the core "Feature Space Cannibalization" problem. We consider the Manifold Capacity $V_{total}$ of the embedding space $\mathbb{S}^{d-1}$ as finite. If seen classes occupy overlapping or expansive regions, $V_{unseen} \to 0$.

To solve this, we propose Hypersphere Energy Regularization using an Angular Margin mechanism. We reinterpret the classification weight vector $\mathbf{w}_c$ as the prototype center for class $c$. The probability of a sample $\mathbf{z}$ belonging to class $y$ is governed by the angular distance $\theta_y = \arccos(\mathbf{w}_y^\top \mathbf{z})$.

To strictly bound the volume $\text{Vol}(\Omega_y)$, we introduce an additive angular penalty $m_{cap}$ (capacity margin):

$$\phi(\theta_y) = \cos\theta_y - m_{cap}. \tag{14}$$

The Capacity Reservation Loss is defined as:

$$\mathcal{L}_{cap} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{s(\cos\theta_{y_i} - m_{cap})}}{e^{s(\cos\theta_{y_i} - m_{cap})} + \sum_{k \neq y_i} e^{s\cos\theta_k}}. \tag{15}$$

where $s$ is the logit scale and $m_{cap}$ is the additive cosine margin. Rather than introducing a new angular-margin loss, we adopt AM-Softmax as an angular compactness regularizer. Our contribution lies in its use and interpretation in the ZS-SBIR setting: the objective encourages embeddings from each seen class to concentrate around their normalized classifier prototype.

Theoretical Insight on Capacity: By enforcing $\cos(\theta_y + m_{cap})$, we force the decision boundary of class $c$ to contract. The acceptance region $\Omega_c$ for class $c$ becomes:

$$\Omega_c = \{\mathbf{z} \in \mathbb{S}^{d-1} \mid \theta_{\mathbf{z},\mathbf{w}_c} \leq \theta_{limit} - m_{cap}\}. \tag{16}$$

As $m_{cap}$ increases, the total volume occupied by seen classes shrinks significantly:

$$\sum_{c \in \mathcal{C}_{seen}} \text{Vol}(\Omega_c) \ll \text{Vol}(\mathbb{S}^{d-1}). \tag{17}$$

This inequality guarantees a non-zero Void Volume ($V_{void} = V_{total} - \sum V_{seen} > 0$), which serves as the "reserved capacity" for mapping unseen categories without overlap. To further stabilize the prototypes, we add a Center Regularization term:

$$\mathcal{L}_{center} = \frac{1}{2} \sum_{i=1}^{N} ||\mathbf{z}_i - \mathbf{w}_{y_i}||_2^2 \tag{18}$$

### 4.5 JOINT OPTIMIZATION OBJECTIVE

The final objective function integrates global anchoring, local rectification, and capacity reservation. We treat this as a multi-objective optimization problem:

$$\mathcal{L}_{total} = \mathcal{L}_{align} + \lambda_{rect}\mathcal{L}_{rect} + \lambda_{cap}(\mathcal{L}_{cap} + \gamma\mathcal{L}_{center}) \tag{19}$$

where $\lambda_{rect}$ and $\lambda_{cap}$ are Lagrangian multipliers balancing the geometric constraints. $\mathcal{L}_{align}$ ensures Global Isometry, $\mathcal{L}_{rect}$ ensures Local Separability, and $\mathcal{L}_{cap}$ ensures Latent Capacity Reservation. By minimizing $\mathcal{L}_{total}$, GADP effectively reconstructs the embedding topology, achieving a compact, shape-biased manifold robust to zero-shot queries.

## 5 EXPERIMENTS

### 5.1 EXPERIMENTAL SETUP

#### 5.1.1 DATASETS SETTINGS

We evaluate GADP on Sketchy Ext., TU-Berlin Ext., and QuickDraw Ext. Sketchy Ext. contains 75,481 sketches and 73,002 images from 125 categories (Sangkloy et al., 2016; Liu et al., 2017). We evaluate both standard zero-shot protocols: Split1 uses 25 unseen categories following Shen et al. (2018), whereas Split2 uses 21 unseen categories that do not overlap with ImageNet following Yelamarthi et al. (2018).

TU-Berlin Ext. contains 20,000 sketches and 204,070 images from 250 categories (Eitz et al., 2012; Liu et al., 2017). Following Shen et al. (2018), we use 220 categories for training and 30 categories for testing.

QuickDraw Ext. contains 330,000 sketches and 204,000 images from 110 categories (Ha & Eck, 2018). Following Dey et al. (2019), we use 80 categories for training and the remaining 30 categories for testing.

#### 5.1.2 IMPLEMENTATION DETAILS

Our method is implemented with PyTorch (Paszke et al., 2017), and all experiments run on an NVIDIA GeForce RTX 5060Ti GPU. We use the pre-trained CLIP ViT-L/14 as sketch and image encoder. The CLIP backbone remains frozen; the learning rate is set to $1\mathrm{e}{-6}$ for LayerNorm parameters and $1\mathrm{e}{-4}$ for prompts. Our framework adopts learnable prompts with dimension 768, matching CLIP's feature size. Two class-agnostic modality-specific prompt banks (one for sketches, one for images) each contain $K = 3$ trainable tokens shared by all seen and unseen categories. No category labels or dedicated unseen-class prompts are needed during inference. GADP follows a Parameter-Efficient Tuning (PET) paradigm. The 300M+ ViT-L/14 backbone is fully frozen; only lightweight prompts and LayerNorm layers are optimized. Trainable parameters account for less than 0.1% of the full model, cutting computation and mitigating overfitting on sparse sketch data. We fix prompt length $K = 3$ with one shallow insertion across all datasets to maintain identical prompt architecture. Our ablations analyze interactions among the three loss terms under this fixed prompt setup. We adopt official standard class splits and set random seed 42 for reproducibility. All reported results are obtained from single deterministic runs.

#### 5.1.3 EVALUATION PROTOCOL

We employ the commonly used evaluation protocol in most studies to assess our model performance. For the Tu-Berlin Ext. and Sketchy Ext. Split1 datasets, we present the average of mean Average Precision (mAP@all) and Precision (Prec@100). For the Sketchy Ext. Split2 dataset, we offer the average of mean Average Precision (mAP@200) and Precision (Prec@200). For the QuickDraw Ext. dataset, we offer the average of mean Average Precision (mAP@all) and Precision (Prec@200).

Table 1: Performance comparison on four ZS-SBIR evaluation settings. A dash indicates that a result was not reported in the original paper. Best results are shown in bold and second-best results are underlined.

| Methods | Year | Dim. | TU-Berlin Ext. | | Sketchy Ext. Split1 | | Sketchy Ext. Split2 | | QuickDraw Ext. | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | mAP@all | Prec@100 | mAP@all | Prec@100 | mAP@200 | Prec@200 | mAP@all | Prec@200 |
| SAKE (Kodirov et al., 2017) | 2017 | 512 | 0.475 | 0.599 | 0.547 | 0.692 | 0.497 | 0.598 | 0.130 | 0.179 |
| Doodle (Dey et al., 2019) | 2019 | 4096 | 0.109 | – | – | – | 0.461 | 0.370 | 0.075 | 0.068 |
| SEM-PCYC (Dutta & Akata, 2019) | 2019 | 64 | 0.297 | 0.426 | 0.349 | 0.463 | – | – | – | – |
| DSN (Wang et al., 2021b) | 2021 | 512 | 0.481 | 0.586 | 0.583 | 0.704 | – | – | – | – |
| Sketch3T (Sain et al., 2022) | 2022 | 512 | 0.507 | – | – | – | – | 0.624 | – | – |
| NAVE (Wang et al., 2021a) | 2021 | 512 | 0.493 | 0.607 | 0.613 | 0.725 | – | – | – | – |
| SBTKNet (Tursun et al., 2022) | 2022 | 512 | 0.480 | 0.608 | 0.480 | 0.608 | 0.502 | 0.596 | – | – |
| CA (Wang et al., 2023) | 2023 | 512 | 0.482 | 0.594 | 0.590 | 0.713 | – | – | – | – |
| RPKD (Tian et al., 2021) | 2021 | 512 | 0.486 | 0.612 | 0.613 | 0.723 | 0.502 | 0.598 | 0.143 | 0.218 |
| OAN (Zhang et al., 2023) | 2023 | – | 0.500 | 0.617 | 0.617 | 0.737 | – | 0.621 | – | – |
| ARML (Liu et al., 2024) | 2024 | 512 | 0.518 | 0.617 | – | – | 0.532 | 0.637 | – | – |
| TVT (Tian et al., 2022) | 2022 | 384 | 0.484 | 0.662 | 0.648 | 0.796 | 0.531 | 0.628 | 0.149 | 0.293 |
| PSKD (ViT) (Wang et al., 2022) | 2022 | 512 | 0.502 | 0.662 | 0.688 | 0.786 | 0.560 | 0.645 | 0.150 | 0.298 |
| ZSE (RN) (Lin et al., 2023) | 2023 | 512 | 0.542 | 0.657 | 0.698 | 0.797 | 0.525 | 0.624 | 0.145 | 0.216 |
| CLIP-ZS-SBIR (Sain et al., 2023) | 2023 | 512 | 0.651 | 0.732 | – | – | 0.723 | 0.725 | 0.202 | 0.388 |
| STL (Ge et al., 2023) | 2023 | – | 0.402 | 0.498 | 0.530 | 0.581 | 0.634 | 0.538 | – | – |
| ZSE (Ret) (Lin et al., 2023) | 2023 | 512 | 0.569 | 0.637 | 0.736 | 0.808 | 0.504 | 0.602 | 0.142 | 0.202 |
| AMA (Yin et al., 2024) | 2024 | – | 0.429 | 0.592 | 0.548 | 0.684 | 0.491 | 0.585 | 0.112 | 0.192 |
| MCG (Zhou et al., 2024) | 2024 | 768 | 0.668 | 0.773 | 0.817 | 0.875 | <u>0.805</u> | 0.768 | 0.332 | <u>0.436</u> |
| IVT (Zhang et al., 2024) | 2024 | 384 | 0.557 | 0.692 | 0.751 | 0.867 | 0.615 | 0.694 | 0.162 | – |
| MARL (Lyou et al., 2024) | 2024 | 512 | 0.707 | 0.777 | – | – | 0.691 | 0.755 | 0.327 | 0.425 |
| SpLIP (Singha et al., 2024) | 2024 | 512 | <u>0.731</u> | <u>0.782</u> | 0.802 | 0.867 | 0.764 | 0.773 | <u>0.342</u> | **0.446** |
| DMWAN (Su et al., 2025) | 2025 | 512 | 0.623 | 0.696 | 0.787 | 0.852 | 0.572 | 0.659 | 0.170 | 0.254 |
| MFCPKD (Li et al., 2025) | 2025 | 768 | 0.662 | 0.728 | <u>0.847</u> | <u>0.878</u> | 0.736 | <u>0.788</u> | 0.224 | 0.285 |
| **GADP (ours)** | – | 768 | **0.804** | **0.832** | **0.922** | **0.937** | **0.810** | **0.852** | **0.345** | 0.402 |

Table 2: Ablation study (%) of the geometric components in GADP on four ZS-SBIR benchmarks.

| Methods | TU-Berlin Ext. | | Sketchy Ext. Split1 | | Sketchy Ext. Split2 | | QuickDraw Ext. | |
|---|---|---|---|---|---|---|---|---|
| | mAP@all | Prec@100 | mAP@all | Prec@100 | mAP@200 | Prec@200 | mAP@all | Prec@200 |
| Global Isometry | 0.710 | 0.739 | 0.875 | 0.885 | 0.741 | 0.794 | 0.234 | 0.285 |
| Global Isometry + Local Rectification | 0.725 | 0.755 | 0.884 | 0.894 | 0.777 | 0.827 | 0.254 | 0.326 |
| Global Isometry + Capacity Reservation | 0.757 | 0.780 | 0.902 | 0.915 | 0.798 | 0.835 | 0.280 | 0.351 |
| **Global Isometry + Local Rectification + Capacity Reservation** | **0.804** | **0.832** | **0.922** | **0.937** | **0.810** | **0.852** | **0.345** | **0.402** |

### 5.2 Comparative Analysis: Mitigating the Generalization Paradox

We compare GADP with state-of-the-art methods across three datasets and four evaluation settings, with results summarized in Table 1. The consistent performance gains indicate that GADP effectively mitigates Feature Space Cannibalization under the zero-shot setting.

Vs. Generative Methods. Generative approaches such as CycleGAN (Zhu et al., 2020) and CAAE (Li et al., 2023) rely on pixel-level translation, which often introduces topological noise and distorts the structural geometry of sketches, leading to unstable decision boundaries—particularly on highly abstract datasets like QuickDraw Ext. In contrast, GADP avoids generation and adapts sketch and image distributions directly within the CLIP manifold via Dual-Prompting, preserving intrinsic topology and yielding superior zero-shot performance.

Vs. Discriminative Methods. Metric learning and fine-tuning based approaches (*e.g.*, Triplet (Schroff et al., 2015), SA-KD (Liu et al., 2021b), ZSE (Lin et al., 2023)) emphasize local distance optimization but ignore the volumetric occupation of seen classes, resulting in capacity exhaustion on the embedding hypersphere. By incorporating Hypersphere Energy Regularization, GADP explicitly compacts seen categories, reserving latent capacity for unseen classes and achieving strong performance across the evaluated settings.

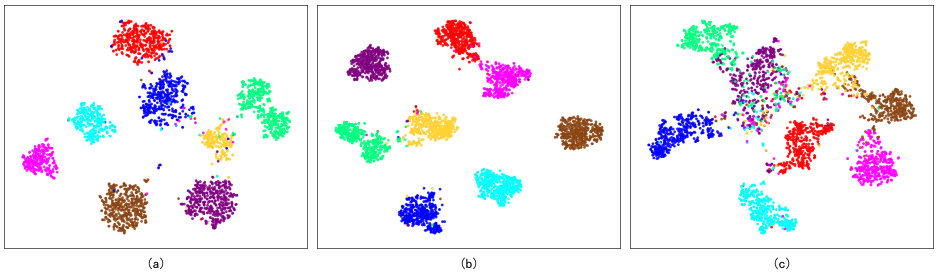



Figure 3: t-SNE visualizations of sketch and photo embeddings on three datasets. We randomly sample examples from eight test categories, with colors denoting categories: (a) TU-Berlin Ext., (b) Sketchy Ext. Split2, and (c) QuickDraw Ext.

### 5.3 Ablation Study: Verifying the Geometric Hypothesis

Since the prompt architecture is fixed, Table 2 isolates the explicit geometric constraints ($\mathcal{L}_{\text{align}}$, $\mathcal{L}_{\text{rect}}$, and $\mathcal{L}_{\text{center}}$) to validate our Manifold Topology Optimization theory ($\mathcal{L}_{\text{cap}}$ is implicitly optimized via CLIP alignment).

- **Global Isometry ($\mathcal{L}_{\text{align}}$):** Establishes basic cross-modal alignment but yields limited performance, proving global alignment alone cannot resolve manifold Cannibalization.
- **Local Rectification ($+\mathcal{L}_{\text{rect}}$):** Rectifies local curvature to push hard negatives beyond the margin, effectively reducing Boundary Ambiguity as reflected by Prec@100 gains.
- **Capacity Reservation ($+\mathcal{L}_{\text{center}}$):** Yields substantial mAP gains by imposing angular margin constraints to shrink intra-class variance. This confirms our core hypothesis: compacting seen classes is essential for reserving latent volume for unseen categories.

### 5.4 Visualizing Manifold Capacity (t-SNE)

We visualize the embedding topology of unseen classes in Figure 3. A striking difference from standard baselines is the Intra-Class Compactness. As shown in Fig. 3(c) (QuickDraw), even for highly abstract sketches, the embeddings form tight, needle-like clusters rather than loose, sprawling clouds.

Crucially, the clear separation between these clusters signifies the existence of "Void Latent Volume"—the empty space on the hypersphere that was successfully reserved during training. This visualization provides direct geometric evidence that GADP effectively prevents seen classes from cannibalizing the open-world capacity.

## 6 Conclusion

In this paper, we reformulate ZS-SBIR as a Manifold Topology Optimization problem to counter Feature Space Cannibalization, where seen categories exhaust finite embedding capacity. To this end, we propose Geometry-Aware Dual-Prompting (GADP), which integrates modality-adaptive contextual tokens with a hypersphere capacity reservation mechanism. By compacting the angular footprint of seen classes through margin-based regularization, GADP explicitly preserves latent capacity for open-world generalization. We further show that manifold capacity is a quantifiable resource in embedding learning and introduce a "reserve-by-compacting" paradigm to mitigate capacity collapse. Beyond ZS-SBIR, this principle generalizes to open-world data mining scenarios such as long-tail recommendation and zero-shot cross-lingual retrieval. Extensive experiments validate that zero-shot generalization is fundamentally governed by the geometric compactness of seen-class distributions, offering a new topological perspective for open-set learning.

### AI USE STATEMENT

The authors used generative AI tools solely for language polishing, improving manuscript organization and clarity, bibliographic verification, and assisting with LaTeX formatting. These tools were not used to design the methodology, conduct experiments, generate or analyze numerical results, or formulate scientific conclusions. The authors carefully reviewed and verified all manuscript content and take full responsibility for the final version.

### REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the problem formulation, model architecture, and optimization objectives. Section 5 describes the datasets, class splits, implementation details, evaluation protocols, and ablation settings. Appendix A provides qualitative failure-case analysis, and Appendix B presents Grad-CAM visualizations. The fixed class splits, prompt configuration, and random seed are reported to facilitate reproduction of the experiments.

### ETHICS STATEMENT

This work uses existing publicly available datasets and pretrained models. No new human-subject study or personally identifiable information was collected. We follow the licenses and intended research uses of the datasets and pretrained models. Cross-modal retrieval systems may inherit biases present in their training data and may perform unevenly across object categories or drawing styles. The reported results should therefore not be interpreted as guarantees of fairness or reliability in safety-critical applications.

## REFERENCES

Saptarshi Dey, Pau Riba, Arijit 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*, pp. 2179–2188, 2019.

Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, pp. 5089–5098, 2019.

Titir Dutta, Anurag Singh, and Soma Biswas. Styleguide: zero-shot sketch-based image retrieval using style-guided image generation. *IEEE Transactions on Multimedia*, 23:2833–2842, 2020.

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

Ce Ge, Jingyu Wang, Qi Qi, Haifeng Sun, Tong Xu, and Jianxin Liao. Semi-transductive learning for generalized zero-shot sketch-based image retrieval. In *AAAI*, volume 37, pp. 7678–7686, 2023.

David Ha and Douglas Eck. A neural representation of sketch drawings. In *International Conference on Learning Representations (ICLR)*, 2018.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, pp. 3174–3183, 2017.

C.-X. Li, D. Zhang, Z. Hu, and X.-J. Wu. Modality fused class-proxy with knowledge distillation for zero-shot sketch-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(6):6158–6169, 2025. doi: 10.1109/TCSVT.2025.3530248.

Chenxi Li, Dong Zhang, and Xinbo Gao. Enhanced conditional adversarial autoencoder for cross-modal sketch-image alignment. *Pattern Recognition Letters*, 169:112–119, 2023. doi: 10.1016/j.patrec.2023.03.021.

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 *CVPR*, pp. 23349–23358, 2023.

Kaiyi Lin, Xing Xu, Lianli Gao, Zheng Wang, and Heng Tao Shen. Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval. In *AAAI*, volume 34, pp. 11515–11522, 2020.

Jian Liu, Feng Wang, and Wei Liu. Angular margin-based losses: A survey and experimental study. *Pattern Recognition*, 118:107919, 2021a. doi: 10.1016/j.patcog.2021.107919.

Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, pp. 2862–2871, 2017.

Qianru Liu, Lingxi Xie, Xin Wang, Qi Zhang, and Qi Tian. Semantic-aware knowledge distillation for sketch-based image retrieval. *IEEE Transactions on Image Processing*, 30:841–854, 2021b.

Yang Liu, Yuhao Dang, Xinbo Gao, Jungong Han, and Ling Shao. Zero-shot sketch-based image retrieval via adaptive relation-aware metric learning. *Pattern Recognition*, 152:110452, 2024. ISSN 0031-3203.

Eunyi Lyou, Doyeon Lee, Jooeun Kim, and Joonseok Lee. Modality-aware representation learning for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5646–5655, January 2024.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

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*, pp. 8748–8763. PmLR, 2021.

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 *CVPR*, pp. 7462–7471, 2022.

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, 2023. URL https://arxiv.org/abs/2303.13440.

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.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, pp. 3598–3607, 2018.

Mainak Singha, Ankit Jha, Divyam Gupta, Pranav Singla, and Biplab Banerjee. Elevating all zero-shot sketch-based image retrieval through multimodal prompt learning, 2024. URL https://arxiv.org/abs/2407.04207.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, volume 28, 2015.

Hanwen Su, Ge Song, Jiyan Wang, and Yuanbo Zhu. Dynamic multi-level weighted alignment network for zero-shot sketch-based image retrieval. *arXiv preprint arXiv:2511.00925*, 2025.

Jialin Tian, Xing Xu, Zheng Wang, Fumin Shen, and Xin Liu. Relationship-preserving knowledge distillation for zero-shot sketch based image retrieval. In *ACM MM*, pp. 5473–5481, 2021.

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 *AAAI*, volume 36, pp. 2370–2378, 2022.

Osman Tursun, Simon Denman, Sridha Sridharan, Ethan Goan, and Clinton Fookes. An efficient framework for zero-shot sketch-based image retrieval. *Pattern Recognition*, 126:108528, 2022.

Feng Wang, Jian Liu, Haijun Liu, Jian Cheng, and Wei Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. doi: 10.1109/LSP.2018.2822810.

Kai Wang, Yifan Wang, Xing Xu, Xin Liu, Weihua Ou, and Huimin Lu. Prototype-based selective knowledge distillation for zero-shot sketch based image retrieval. In *ACM MM*, pp. 601–609, 2022.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Wenjie Wang, Yufeng Shi, Shiming Chen, Qinmu Peng, Feng Zheng, and Xinge You. Norm-guided adaptive visual embedding for zero-shot sketch-based image retrieval. In *IJCAI*, pp. 1106–1112, 2021a.

Xiaohan Wang, Dezhong Peng, Peng Hu, Yihong Gong, and Yurong Chen. Cross-domain alignment for zero-shot sketch-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Zhipeng Wang, Hao Wang, Jiexi Yan, Aming Wu, and Cheng Deng. Domain-smoothing network for zero-shot sketch-based image retrieval. *arXiv preprint arXiv:2106.11841*, 2021b.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7959–7971, 2022.

F. Yang, N. A. Ismail, Y. Y. Pang, V. R. Kebande, A. Al-Dhaqm, and T. W. Koh. A systematic literature review of deep learning approaches for sketch-based image retrieval: Datasets, metrics, and future directions. *IEEE Access*, 12:14847–14869, 2024.

Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, pp. 300–317, 2018.

Z. Yin, J. Yan, C. Xu, and C. Deng. Asymmetric mutual alignment for unsupervised zero-shot sketch-based image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16504–16512, 2024.

Haoxiang Zhang, He Jiang, Ziqiang Wang, and Deqiang Cheng. Ontology-aware network for zero-shot sketch-based image retrieval. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Haoxiang Zhang, Deqiang Cheng, Qiqi Kou, Mujtaba Asad, and He Jiang. Indicative vision transformer for end-to-end zero-shot sketch-based image retrieval. *Advanced Engineering Informatics*, 60:102398, 2024. ISSN 1474-0346.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, July 2022a. ISSN 1573-1405. doi: 10.1007/s11263-022-01653-1. URL http://dx.doi.org/10.1007/s11263-022-01653-1.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models, 2022b. URL https://arxiv.org/abs/2203.05557.

Yanghong Zhou, Dawei Liu, and PY Mok. Zero-shot sketch based image retrieval via modality capacity guidance. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 1780–1787, 2024.

Jiawen Zhu, Xing Xu, Fumin Shen, Roy Ka-Wei Lee, Zheng Wang, and Heng Tao Shen. Ocean: A dual learning approach for generalized zero-shot sketch-based image retrieval. In *ICME*, pp. 1–6. IEEE, 2020.

## A Failure Case Analysis (Topology Confusion)

Analyzing the failure cases (red crosses in Fig. 4,5 and 6) reveals that errors are mostly "Topologically Isomorphic". For instance, a "cat" sketch is retrieved as a "tiger" (Fig. 4). This is not a semantic failure but a geometric one: the two share a near-identical structural topology. This suggests that while GADP successfully aligns the manifold, future work could focus on fine-grained topological disentanglement for geometrically similar categories.

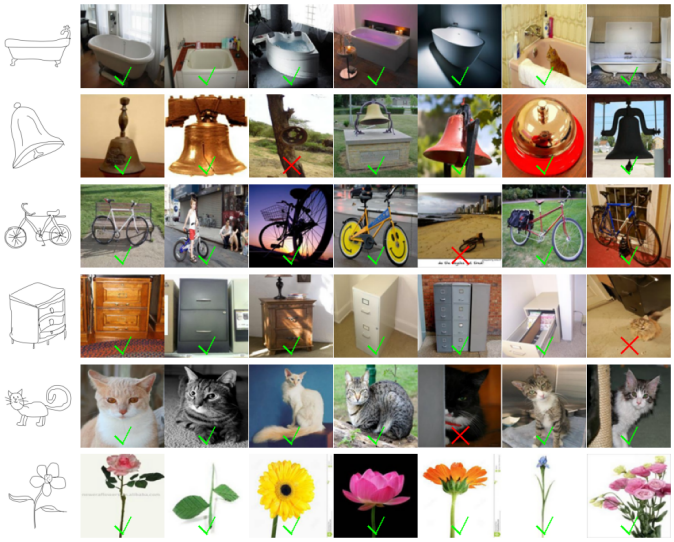


Figure 4: The top 7 real images retrieved through sketch-based search in the TU-Berlin Ext. dataset. All samples are from unseen categories. We use ticks and crosses to denote correct and incorrect retrieval results, respectively. Distinguishing between similar objects during retrieval poses a challenge for our model.

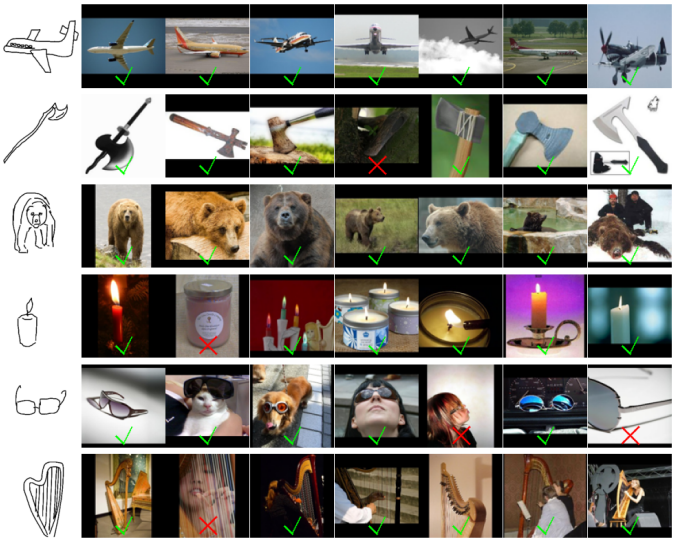


Figure 5: The top 7 real images retrieved through sketch-based search in the Sketchy Ext. Split2 dataset. All samples are from unseen categories. We use ticks and crosses to denote correct and incorrect retrieval results, respectively. Distinguishing between similar objects during retrieval poses a challenge for our model.

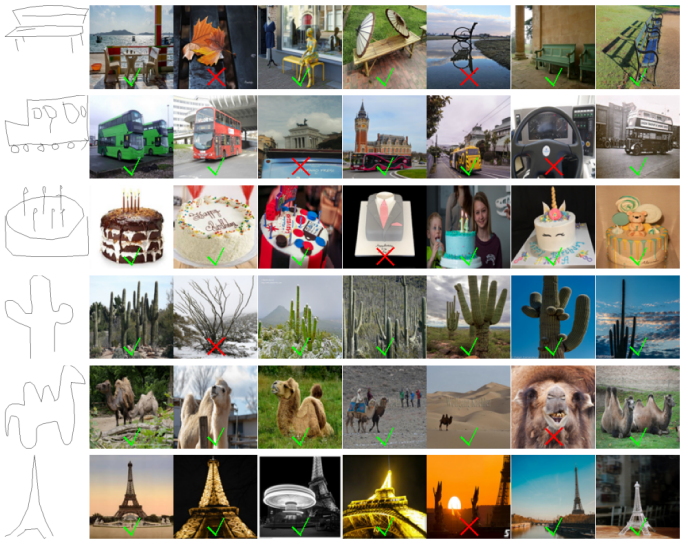


Figure 6: The top 7 real images retrieved through sketch-based search in the QuickDraw Ext. dataset. All samples are from unseen categories. We use ticks and crosses to denote correct and incorrect retrieval results, respectively. Distinguishing between similar objects during retrieval poses a challenge for our model.

## B VISUALIZING STRUCTURAL DECOUPLING (GRAD-CAM)

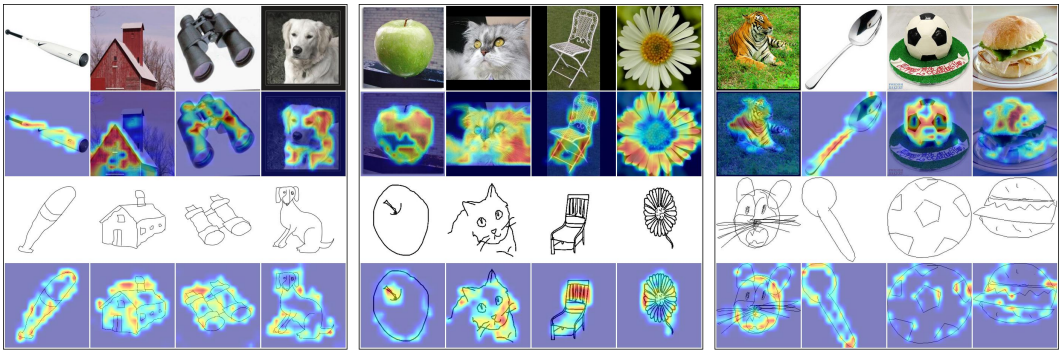



Figure 7: Grad-CAM visualizations of feature attention on (a) TU-Berlin Ext., (b) Sketchy Ext. Split2, and (c) QuickDraw Ext.

Figure 7 illustrates the attention maps. The model consistently focuses on topological structures (*e.g.*, the circular contour of the "ball" or the facial geometry of the "cat") while ignoring textural noise in images. This confirms that our Modality-Adaptive Contextual Tokens function as effective "Distribution Anchors". They act as a high-pass filter, decoupling the essential structural topology from modality-specific noise, ensuring that the alignment occurs on a shape-biased semantic manifold.