Homology Consistency Constrained Efficient Tuning for Vision-Language Models

Huatian Zhang, Lei Zhang, Yongdong Zhang, Zhendong Mao*
University of Science and Technology of China
huatianzhang@mail.ustc.edu.cn, {leizh23,zhyd73,zdmao}@ustc.edu.cn

Abstract

Efficient transfer learning has shown remarkable performance in tuning large-scale vision-language models (VLMs) toward downstream tasks with limited data resources. The key challenge of efficient transfer lies in adjusting image-text alignment to be task-specific while preserving pre-trained general knowledge. However, existing methods adjust image-text alignment merely on a set of observed samples, *e.g.*, data set and external knowledge base, which cannot guarantee to keep the correspondence of general concepts between image and text latent manifolds without being disrupted and thereby a weak generalization of the adjusted alignment. In this work, we propose a Homology Consistency (HC) constraint for efficient transfer on VLMs, which explicitly constrains the correspondence of image and text latent manifolds through structural equivalence based on persistent homology in downstream tuning. Specifically, we build simplicial complex on the top of data to mimic the topology of latent manifolds, then track the persistence of the homology classes of topological features across multiple scales, and guide the directions of persistence tracks in image and text manifolds to coincide each other, with a deviating perturbation additionally. For practical application, we tailor the implementation of our proposed HC constraint for two main paradigms of adapter tuning. Extensive experiments on few-shot learning over 11 datasets and domain generalization demonstrate the effectiveness and robustness of our method.

1 Introduction

Large-scale vision-language models (VLMs) such as CLIP [1] and ALIGN [2] trained on web-scale data have learned broad visual concepts and demonstrated promising generalization capability on a wide range of downstream tasks, such as classification [3, 4], detection [5, 6] and segmentation [7, 8]. In transferring to downstream tasks with limited data resources, the conventional full fine-tuning on VLMs often forgets the general knowledge learned in pre-training and falls into overfitting. To mitigate this, how to efficiently transfer the knowledge from pre-trained VLMs to downstream tasks in a low-data regime has been intensively studied.

To possess both task-specific knowledge exploration and general knowledge preservation, efficient transfer learning proposes to adapt VLMs to fit downstream tasks by tuning a few parameters, which mainly exists in two paradigms: prompt tuning and adapter tuning. Prompt tuning methods adapt VLMs toward downstream tasks by introducing learnable prompts on the input side. Topics in this branch include the configuration of learnable prompts [9–13] and injecting semantic priors, such as external knowledge [14, 15], category distribution [16] or visual diversity [17, 18], in tuning. As a promising alternative, adapter tuning inserts a learnable lightweight adapter into the frozen pre-trained VLMs on the output side and the insertion manner allows modifying flexibly. By residual blending, CLIP-Adapter [19] tuned the image or text embeddings by appending a learnable bottleneck layer

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

to the frozen encoder. TaskRes [20] added learnable parameters as prior-independent residuals to text embeddings. Another line is based on key-value cache model. Tip-Adapter [21] proposed to construct an adapter with training images as keys and one-hot label encodings as values for image recognition. Some work further enhanced the cache model by discriminative prior refinement [22] or knowledge augmentation [23]. For efficient transfer, existing methods mainly focus on how to configure tunable parameters or the leverage of external prior knowledge in tuning.

The essence of task-specific tuning on VLMs is to adjust the semantic alignment of image and text latent manifolds to fit downstream tasks. In this view, efficient transfer learning aims to establish new semantic alignments while keeping the correspondence of pre-trained general concepts between image and text latent manifolds from being corrupted. This correspondence stems from the equivalence of semantics between the two manifolds. However, existing methods adjust image-text alignment toward downstream tasks on a set of observed samples from the latent manifolds. The discrete samples are incapable of adequately capturing the underlying structure of manifolds. The lack of perspective on latent manifolds in alignment adjusting risks devolving the desired manifold equivalence into localized closeness on the observed data, particularly in a low-data regime, which causes the unguaranteed generalization of adjusted alignment beyond the data samples.

To this end, we propose to explicitly constrain the equivalence of image and text latent manifolds in transferring VLMs toward downstream tasks. We study the structure of the image and text latent manifolds from the lens of topological data analysis [24–26]. Topology encodes the connectivity of a space to describe its underlying structure, and the preservation of topology between spaces is fundamental for their structural equivalence. Topological data analysis employs homology groups to quantify the topological features of manifold structure, such as connected components, loops, voids, and higher-dimensional holes, as homology classes and tracks the survival of the topological features across multiple scales via persistent homology [26] to capture their size and position, which summaries the global shape of manifold. These topology insights provide an avenue to achieve a structural equivalence of image and text latent manifolds.

In this work, we propose a Homology Consistency (HC) constraint for efficient transfer on VLMs, which constrains the structural equivalence of image and text latent manifolds based on persistent homology in downstream tuning. Given image-text data samples, we construct simplicial complex on the top of data to mimic the topological structure of latent manifolds, and induce a nested sequence of subcomplexes called filtration. Through the filtration, we capture the persistence of homology classes from their appearance to non-existence and guide the homology persistences of image and text manifolds to be consistent. Specifically, we locate the births and deaths of homology classes in latent manifolds and track these homology persistences, then guide the directions of persistence tracks in image and text manifolds to coincide each other so as to achieve a homology-level structural equivalence. Additionally, we apply a deviating perturbation to persistence-related text samples to encourage their respective semantically related images to be distributed uniformly relative to them in embedding, in order to enhance the generalization of the track coincidence in adjusting image-text alignment. Further, we tailor the implementation of the proposed HC constraint for the main residual blending and key-value cache based paradigms of adapter tuning. Extensive experiments on few-shot learning over 11 benchmarks and domain generalization demonstrate the effectiveness and robustness of HC constraint in efficient transfer learning on VLMs.

Our main contributions are summarized as follows:

- We propose to explicitly constrain the structural equivalence of image and text latent manifolds in efficient transfer on VLMs, to improve the generalization of downstream image-text alignment adjusting beyond data samples in a low-data regime.
- We propose a theoretically well-founded homology consistency (HC) constraint based on persistent homology for efficient transfer on VLMs. We coincide the persistences of homology classes of topological features between image and text manifolds, and apply a deviating perturbation to generalize the persistence coincidence to unseen data.
- We tailor the implementation of the proposed HC constraint for the two main paradigms of adapter tuning respectively, showing the transferability of our method.
- We evaluate the proposed HC constraint on few-shot classification over 11 popular benchmarks. The extensive experiments demonstrate that HC constraint can boost the performance of baselines significantly and achieve state-of-the-art.

2 Related Work

Efficient Transfer Learning. To better transfer VLMs to downstream tasks especially with limited target domain data, a lot of research on efficient transfer learning has been done, which mainly exists in two types, prompt tuning and adapter tuning. Advanced than manual prompt that demands domain expertise to develop suitable format, prompt tuning methods design learnable prompts to adapt VLMs on downstream data. As a pioneer work, CoOp [9] for the first time composed prompts by concatenating text category embedding and learnable context vectors. The learnable prompts can be configured in text input [9, 10], image input [11] or jointly both [12, 13]. A line of work focuses on injecting semantic priors, such as category-related external knowledge [14, 15], category embedding distribution [16] and the diversity of visual concepts [17, 18], in prompt tuning. In another branch, adapter tuning methods insert a learnable lightweight adapter module into the frozen pre-trained VLMs and show excellent performance. The adapter architecture and insertion manner allow for flexible modifications. There are two main paradigms: residual blending and key-value cache based. By residual blending, CLIP-Adapter [19] appended learnable bottleneck layer to frozen encoder to tune embeddings. TaskRes [20] added a set of prior-independent parameters to frozen text category embeddings to obtain an image classifier. GraphAdapter [27] proposed to learn downstream knowledge with inter-class relationship of image and text samples. Based on the key-value cache model, Tip-Adapter [21] constructed an adapter with training images and label encodings as key-value to recognize query images. APE [22] refined the cache model by visual discrimination priors. CaFo [23] cascaded diverse external knowledge from DINO [28], DALL-E [29], and GPT-3 [30] to assist recognition. However, although existing methods have attained remarkable achievements in VLMs transfer, they adjust the semantic alignment [31–33] of image and text latent manifolds toward downstream tasks merely on observed samples, *e.g.*, data sets and external knowledge bases, and lack insight into underlying manifold structure, which may cause unguaranteed generalization beyond the data samples. In this work, we propose to explicitly constrain the structural equivalence of image and text latent manifolds in downstream tuning to facilitate the transfer of VLMs.

Topological Data Analysis in Machine Learning. The area of topological data analysis [26] infers the topological structure of data spaces using algebraic tools such as persistent homology, and has been applied in many fields of machine learning, *e.g.*, image segmentation [34–37], graph machine learning [38–40], molecular representation [41, 42], point cloud analysis [43, 44], etc. For instance, in image segmentation, [35] proposed to drive the segmentations to contain the specified topological features without requiring ground-truth labels. [36] used discrete Morse theory and persistent homology to learn the structural representation of images for fine-scale structure segmentation. In graph machine learning, [39] integrated vertex- and edge-level topological features into message-passing graph neural networks to boost their expressive power. In point cloud analysis, [44] developed a learnable filtration on point clouds to obtain adaptive topological features for given tasks. Besides, [45] preserved the topological structures of input space into latent space of autoencoders by aligning topologically relevant distances. [46] applied representation topology divergence [47] in dimensionality reduction to force closeness on topological structures. How to analyse the structure of data spaces for semantic alignment in vision-language tasks through topology remains under-explored in the literature. In this work, we characterize and align the structure of the image and text latent manifolds by means of persistent homology in efficient transfer learning on VLMs.

3 Methodology

3.1 Preliminaries

3.1.1 Contrastive Language-Image Pre-training (CLIP)

As a representative VLM, CLIP [1] is trained on massive image-text pairs and shows promising zero-shot performance on downstream tasks. CLIP adopts two separate encoders to embed images and texts into latent manifolds, and aligns the bi-modal embeddings by contrastive learning that forces paired image and text closer and unpaired ones away. For the transfer to image classification with N classes, CLIP obtains the class embedding w_i by feeding the prompt templates, *e.g.*, "A photo of a [CLASS]", filled with class name c_i into text encoder, and the probability that an image x belongs to category c_i can be formulated as $p(y=c_i|x) = \exp\left(x^\top w_i/\tau\right) / \sum_{j=1}^{N} \exp\left(x^\top w_j/\tau\right)$, where the embeddings are l_2-normalized and τ denotes a temperature hyper-parameter.

3.1.2 Persistent Homology

Discrete data points contain only observations from the latent space in which they reside and do not have interesting topology. To peek at the topological structure of latent space, we mimic its connectivity by constructing a simplicial complex with the data points as vertices. As a basic element, simplex σ with dimension p is the convex hull of a set of $p+1$ affinely independent points $(x_0, \ldots, x_p)$. Given a finite data point set X in metric space (M, d) and a threshold $a > 0$, the commonly used Vietoris-Rips (Rips in short) complex is defined as:

$$K_a(X) = \{\sigma \subset X \mid d(x_i, x_j) < a, \forall x_i, x_j \in \sigma\}, \tag{1}$$

which is fully determined by the pairwise distances of X. The dimension of K_a is the maximum dimension of any simplex within it. The formal sums of p-simplices added with $\mathbb{Z}_2$-additions form a chain group $\mathrm{C}_p(K_a)$ (C_p for brevity). Then, define a boundary operator ∂_p on p-simplex σ as a map that sends σ to the $(p-1)$-chain consisting of σ's $(p-1)$-faces referred as σ 's boundary. Applying ∂_p to the chain groups can obtain a sequence of homomorphisms: $\mathrm{C}_k \xrightarrow{\partial_k} \mathrm{C}_{k-1} \cdots \mathrm{C}_1 \xrightarrow{\partial_1} \mathrm{C}_0$. All p-chains whose boundaries are empty form a cycle group Z_p, which is the kernel of ∂_p. The image of boundary operator ∂_{p+1} on C_{p+1} forms a boundary group B_p. Further, taking the quotient of the Z_p with B_p, the p-th homology group $\mathrm{H}_p = \mathrm{Z}_p/\mathrm{B}_p$ classifies the p-cycles in Z_p by collecting those cycles that differ by a boundary into the same homology class. In a topological view, the rank of homology group H_p captures the number of p-dimensional holes in space.

Persistent homology offers a way to compute the quantified summary of topological structures of the latent space from sampled data. For the data point set X in space (M, d), we define a function $f : M \to \mathbb{R}, f((x_0, \ldots, x_p)) = \max_{i,j \in \{0,\ldots,p\}} f((x_i, x_j))$ on simplices. Then, given a sequence of thresholds $a_1 \leq a_2 \leq \ldots, \leq a_n$, the growing sublevel sets $f^{-1}(-\infty, a]$ at these values give rise to a nested sequence of subcomplexes, $K_{a_1} \subseteq K_{a_2} \subseteq \cdots \subseteq K_{a_n}$, called a filtration $\mathcal{F}$, as shown in Fig. 1. The inclusions in $\mathcal{F}$ induce:

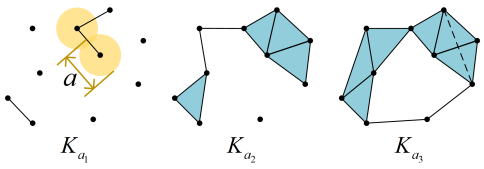


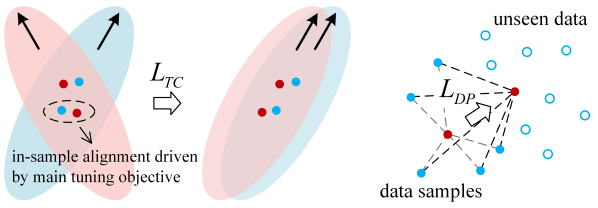
Figure 1: The sublevel set filtration on a nested family of Rips complexes.

$$0 = \mathrm{H}_p(K_0) \to \cdots \to \mathrm{H}_p(K_i) \xrightarrow{h_p^{i,j}} \cdots \to \mathrm{H}_p(K_j) \to \cdots \to \mathrm{H}_p(K_n) = \mathrm{H}_p(K), \tag{2}$$

where the images of the homomorphisms $h_p^{i,j}$ are persistent homology groups $\mathrm{H}_p^{i,j}$. A non-trivial homology class $\epsilon \in \mathrm{H}_p(K_a)$ is born at K_i, if $\epsilon \in \mathrm{H}_p^{i,a}$ but $\epsilon \notin \mathrm{H}_p^{i-1,a}$. Likewise, the homology class ϵ dies entering K_j, if $\epsilon \in \mathrm{H}_p^{a,j-1}$ but $\epsilon \notin \mathrm{H}_p^{a,j}$. The persistence of a homology class is the lifespan from its birth to death. See Appendix A for more details.

3.2 Homology Consistency

Given a set of pre-trained image embeddings X and text embeddings T, we construct a Rips complex $K_{a_M}(X)$ with the maximum pairwise distance a_M of X and further derive a sublevel set filtration, $\mathcal{F}(K)$, as the nested family of subcomplexes $K_{a_0} \subseteq K_{a_1} \subseteq \cdots \subseteq K_{a_M}$ at the increasing scale sequence of pairwise distances $\{a_i\}_{i=0}^M$ where $a_0 = 0$. Then we arrive at p-th persistent homology groups that capture the survival of the homology classes of p-dimensional topological features (*e.g.*, 0-dimension: connected components, 1-dimension: loops, 2-dimension: voids, etc.) and pair the birth and death times of p-th homology classes, following [45]. Since the edge skeleton of Rips complex fully determines all of its simplices,



(a) Persistence Track Coincidence (b) Deviating Perturbation

Figure 2: Schematic illustration of our proposed HC constraint. (a) TC guides the directions of the persistence tracks to coincide each other to establish the alignment of underlying structures beyond the observed samples. (b) DP encourages samples to be uniformly distributed.

under the assumption that pairwise distances of X are unique, every birth-death time pair can be mapped back to the simplices that respectively created and destroyed the unique corresponding homology class. The schematic illustration of HC constraint is shown in Fig. 2.

Persistence Track Coincidence. Let μ be a p-th homology class that is born at a_b and dies at a_d, then the birth simplex β forms μ in K_{a_b} and the death simplex δ causes μ to disappear entering K_{a_d}. For example, if μ were to be 0-dimensional, it emerges at a_0, and δ is the edge that joins it with some other point; if μ were to be 1-dimensional, then β is the edge that forms the loop corresponding to μ in K_{a_b}, and δ is the triangle that incurs the loop to be contractible in K_{a_d}. Since the Rips complex is pairwise distance based, on every p-simplex we have:

$$f\left((x_0, \ldots, x_p)\right) = \max_{i,j \in \{0,\ldots,p\}} f\left((x_i, x_j)\right). \tag{3}$$

For the birth simplex β and death simplex δ, this means they will not be established until their longest edge $(x_{i(\beta)}, x_{j(\beta)})_{\arg\max f(\beta)}$ and $(x_{i(\delta)}, x_{j(\delta)})_{\arg\max f(\delta)}$ appear at time a_b and a_d, which we refer to as birth edge and death edge, respectively. It can be said that the simplex β is completed by the birth edge and δ by the death edge. Such edges mark the creation and destruction of the homology class μ in the image latent manifold. Defining the direction of the edges as from i to j such that $i < j$ in the given vertex sequence of simplices, we track the persistence direction of p-th homology class μ from its birth edge to death edge as:

$$\mathrm{tra}_p(\mu, X) = (x_{i(\delta)}, x_{j(\delta)})_{\arg\max f(\delta)} - (x_{i(\beta)}, x_{j(\beta)})_{\arg\max f(\beta)}. \tag{4}$$

Then we obtain the persistence track of μ in text latent manifold, $\mathrm{tra}_p(\mu, T)$, through determining its birth and death edges directly by taking the corresponding texts of the end-points of μ's birth and death edges in image manifold accordingly. Further, for aligning the structure of image and text latent manifolds, we guide the track coincidence (TC) of p-th homology classes between the two by:

$$L_{TC}(\Gamma_p, X, T) = 1 - \frac{1}{|\Gamma_p|} \sum_{\mu \in \Gamma_p} \varphi(\mathrm{tra}_p(\mu, X), \mathrm{tra}_p(\mu, T)), \tag{5}$$

where Γ_p is the set of p-th homology classes in latent manifolds and φ is cosine similarity.

Deviating Perturbation. In a low-data regime where samples are not sufficient to fully characterize the topology of the latent manifold of interset, the sight of persistence track coinciding is confined on the given limited samples, which hinders the generalization of structural equivalence guided by track coincidence. For all end-point images of birth and death edges of p-th homology classes in Γ_p and their corresponding texts, we drive every text to deviate from its semantically related images in embedding (what we consider here is that multiple images are related to the same category text in classification) without breaking the track coincidence, so as to encourage the text-related images beyond samples in latent manifold to be uniformly distributed around the text.

Specifically, since the persistence tracks of the 0-th homology classes are exactly the death edges (birth edges are 0), their end-point samples are available from the given dataset. We quantify the degree of deviation as the similarity between the orientations of the text's embedding relative to the embeddings of its related images. Then, for 0-th homology classes in Γ_0, we enlarge the deviation between end-point texts and their respective semantically related images by reducing the variation in their relative orientations, by applying a deviating perturbation (DP) as:

$$L_{DP}(\Gamma_0, X, T) = \frac{1}{|X_{\Gamma_0}|} \sum_{x_i \in X_{\Gamma_0}} (1 - \frac{1}{|X_i'|} \sum_{x_i' \in X_i'} \varphi(x_i - (x_i)_T, x_i' - (x_i)_T)), \tag{6}$$

where X_{Γ_0} denotes the set of all end-point images of 0-th homology classes in Γ_0, $(x_i)_T$ denotes x_i's corresponding text, $X_i' \in X_{\Gamma_0}$ denotes a set of images of the same category that are semantically related to x_i and φ is cosine similarity. The deviating perturbation can benefit the structural equivalence to get rid of a biased reconstruction of latent topology caused by insufficient sampling.

Homology Consistency. To constrain the structural equivalence of image and text latent manifolds, we coincide persistence tracks of homology classes along with the deviating perturbation by:

$$L_{HC}(\Gamma, X, T) = \sum_{p=0}^{n} L_{TC}(\Gamma_p, X, T) + \lambda L_{DP}(\Gamma_0, X, T), \tag{7}$$

where λ is the hyper-parameter controlling the perturbation strength. We mainly focus on the 0-th homology classes ($p = 0$), since higher-dimensional classes significantly increase computational cost but bring almost no additional performance benefits in VLMs tuning in practice.

3.3 Implementation

To examine the structural equivalence constrained by our proposed homology consistency in the efficient transfer of VLMs and show its transferability, we tailor the implementation of homology consistency constraint to the two main paradigms of adapter tuning, residual blending based and key-value cache model based, respectively.

Residual blending based tuning methods construct an adapter to produce learnable residuals and blend them with the pre-trained features. As a representative, TaskRes [20] adds prior-independent parameters x as a residual to the pre-trained text embeddings t to form a learnable image classifier $t' = t + \alpha x$, where α is a scaling factor, and updates the classifier by cross-entropy loss L_{CE}.

In this paradigm, given frozen pre-trained image embeddings X and tunable text embeddings T, our method can naturally construct a filtration on X to capture the image persistence tracks and further obtain their text tunable counterparts, and then employ the persistence track coincidence with deviating perturbation through $L_{HC}(\Gamma, X, T)$ in downstream tuning together with L_{CE}.

Key-value cache based tuning methods build adapters via a key-value cache model with pre-trained embeddings of all training images as keys and one-hot encodings of corresponding labels as values. The image keys can be unfreezed as learnable parameters. As a representative, to recognize an image x, Tip-Adapter-F [21] first measures its affinity weights with the cached keys F_{train} by $A = \exp\left(-\beta\left(1 - xF_{\text{train}}^T\right)\right)$, then aggregates the cached values L_{train} with weights A as a prediction AL_{train}, and further combines AL_{train} with the similarity between image x and pre-trained text category embeddings W_c as final classification logit $\alpha AL_{\text{train}} + xW_c^T/\tau_{\text{CLIP}}$ in L_{CE}.

In this paradigm, the cache-based adapter represents the categories as one-hot label encodings L_{train}, that is, the L_{train} is the only textual representation of categories in this adapter. To be comparable to label encodings, we regard the affinity weights A of an image as a sparse visual representation of this image. For implementing homology consistency, we first capture the birth and death edges of image homology classes based on pre-trained image embeddings, then replace end-point images of the edges with their corresponding affinity weights A or label encoding L_{train} and follow Eq. 4 to construct $\text{tra}_p(\mu, A)$ and $\text{tra}_p(\mu, L_{\text{train}})$ analogically. We implement HC as $L_{HC}(\Gamma, A, L_{\text{train}})$ by taking $\text{tra}_p(\mu, A)$ and $\text{tra}_p(\mu, L_{\text{train}})$ as proxies for original image and text tracks.

Optimization. Efficient transfer learning commonly adopts the cross-entropy loss L_{CE} between labels and the predicted class probability to tune learnable parameters, *e.g.*, θ, on downstream classification tasks. We have two ways to integrate the gradient from our proposed HC constraint into parameter tuning. (1) We alter the cross-entropy gradient $\nabla_\theta L_{CE}$ toward the direction of HC by adding the gradient of HC constraint to $\nabla_\theta L_{CE}$ with a constant factor η by: $\nabla_\theta L = \nabla_\theta L_{CE} + \eta \nabla_\theta L_{HC}$. (2) For Tip-Adapter-F, due to its slightly complex hyper-parameter configuration, fixing a constant factor is difficult to control the contribution of each gradient. We adaptively keep the gradients at the same order of magnitude through scaling the gradient of homology consistency constraint based on gradient norm by $\nabla_\theta L = \nabla_\theta L_{CE} + \omega \frac{\|\nabla_\theta L_{CE}\|_2}{\|\nabla_\theta L_{HC}\|_2} \nabla_\theta L_{HC}$ in optimization.

4 Experiments

4.1 Experimental Settings

Datasets. Following previous efficient transfer learning works, we conduct the few-shot learning evaluation on 11 benchmark datasets including Caltech101 [48], DTD [49], EuroSAT [50], FGVCAircraft [51], Flowers102 [52], Food101 [53], ImageNet [54], OxfordPets [55], StanfordCars [56], SUN397 [57] and UCF101 [58]. These datasets cover a wide range of visual recognition on generic objects, fine-grained categories, scenes, actions, etc. We sample 1, 2, 4, 8 and 16 shots per class, respectively, for model training and evaluate on full test sets. In addition, we evaluate the domain generalization performance of our method with the ImageNet [54] as source and its variants ImageNetV2 [59], ImageNet-Sketch [60], ImageNet-A [61] and ImageNet-R [62] as targets.

Implementation Details. In addition to HC, following TaskRes, we augment the original pre-trained text features into ones tuned on downstream tasks in HC*. In implementing HC / HC* on TaskRes (HC-TR / HC*-TR), we set the values of η and λ on different datasets according to the procedure of factor determination in ablation studies. We set the scaling factor α to 1 for all datasets. The training batch size is 256. We employ the Adam optimizer with an initial learning rate of $1e^{-4}$

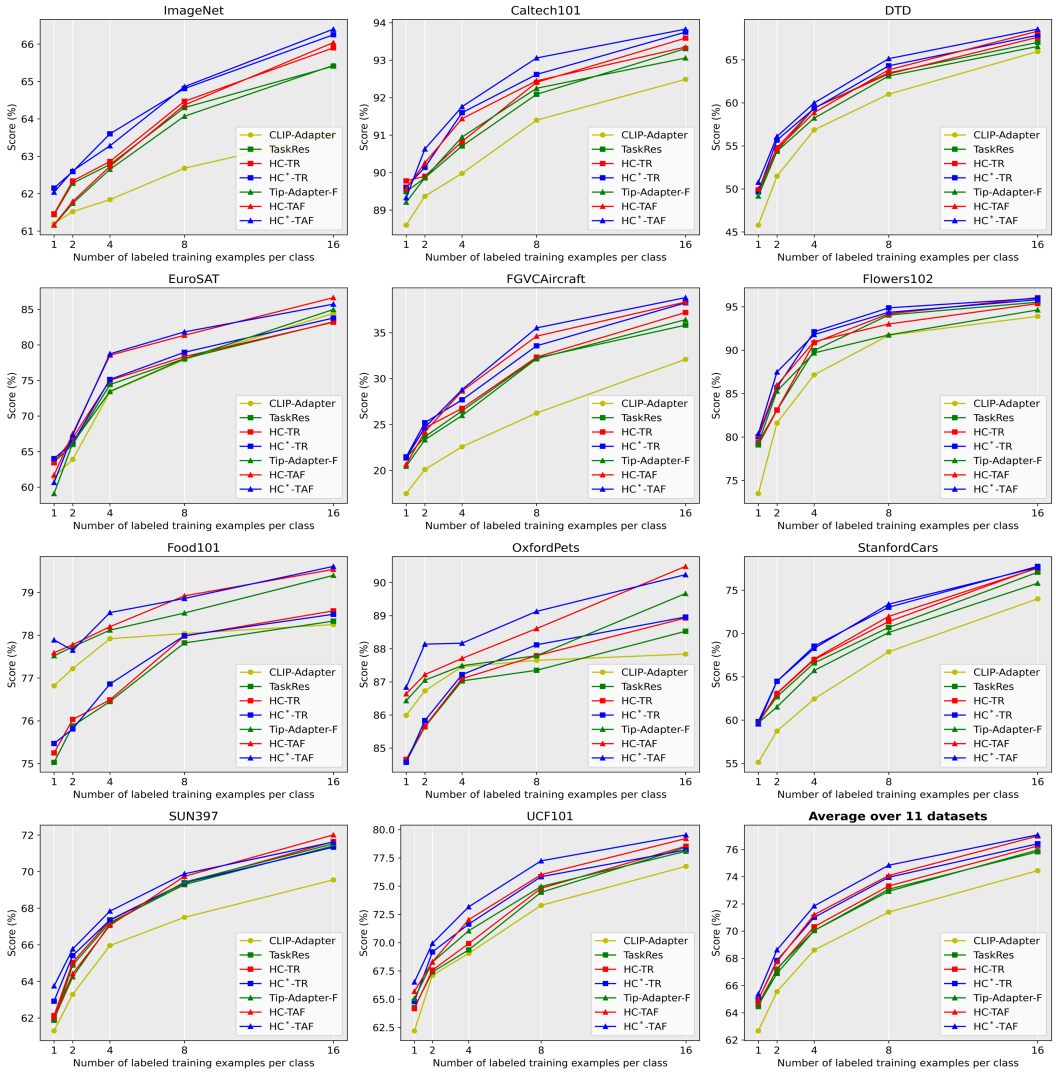


Figure 3: The performance comparison of baselines and our proposed HC and HC* on few-shot learning, including 1-/2-/4-/8-/16-shots on 11 benchmark datasets and the average accuracy. The full numerical results are provided in the Appendix B.

on ImageNet and $1e^{-3}$ on others, and the learning rates decay with cosine learning rate schedule following TaskRes. In implementing HC / HC* on Tip-Adapter-F (HC-TAF / HC*-TAF), ω and λ are also set according to the factor determination procedure in ablation studies. The λ here is significantly larger than the above because the magnitude difference between the $\nabla_\theta L_{TC}$ and $\nabla_\theta L_{DP}$ here is relatively large. The training batch size is 256. Following Tip-Adapter-F, we employ the AdamW optimizer with a cosine annealing scheduler. We set initial learning rate as $1e^{-3}$. All experiments are conducted on a single NVIDIA A40 GPU. Note that the experiments on baseline and with the HC / HC* are conducted in the same setting for fair comparison. The code is publicly available [2].

4.2 Performance Analysis

Few-shot Learning. We validate the effectiveness of our proposed HC constraint on 11 benchmark datasets using representative adapter tuning methods, the residual blending based TaskRes and key-value cache based Tip-Adapter-F, as baselines. The experimental results are shown in Fig. 3, from which it can be observed that our method brings performance improvements for 1 to 16 shots

[2]https://github.com/htzhang-code/HC

consistently. Taking the 16 shots average accuracy as an example, HC-TR exceeds TaskRes by 0.40% and HC-TAF exceeds Tip-Adapter-F by 1.04%. At 16-shot, HC-TAF outperforms Tip-Adapter-F by 0.61% on ImageNet. For the challenging fine-grained classification dataset FGVCAircraft, the proposed HC constraint gains 1.38% and 1.92% on TaskRes and Tip-Adapter-F, respectively. The further improvements of HC*-TR and HC*-TAF yielded by pre-tuned text classifier suggest that the HC-equipped tuning can still significantly benefit from representation enhancement.

The performance comparison of our HC constraint with the state-of-the-art adapter tuning methods on ImageNet is shown in Tab. 1. APE-T performs markedly best at 1 and 2 shots, whereas the models constrained by homology consistency, *e.g.*, HC*-TAF, exceed it and other state-of-the-arts on the 8-/16-shot setting in more sample cases. This arises from the fact that the efficacy of homology consistency in vision-language aligning depends primarily on the modeling capability of the simplicial complex builded on training data towards the topological structure of latent space. The denser the data samples, the more sufficiently they summarize the structure, and the more effective the homology consistency constraint is. It is worth noting that HC boosts baselines to achieve state-of-the-art without introducing any additional training parameters. We respectively take TaskRes and Tip-Adapter-F as representative methods of residual blending based (*e.g.*, CLIP-Adapter, TaskRes, GraphAdapter) and key-value cache based (*e.g.*, Tip-Adapter-F, APE-T) adapter tuning for applying HC constraint. The HC constraint is verified to be effective on these two baselines by extensive experiments, and can theoretically be extended to other VLMs efficient transfer learning methods.

Table 1: The performance comparison of our methods with the state-of-the-art methods on ImageNet.

Method	1-shot	2-shot	4-shot	8-shot	16-shot
Zero-shot CLIP [1]	58.18	58.18	58.18	58.18	58.18
CLIP-Adapter [19]	61.20	61.52	61.84	62.68	63.59
TaskRes [20]	61.44	62.28	62.78	64.30	65.41
Tip-Adapter-F [21]	61.16	61.74	62.65	64.07	65.43
GraphAdapter [27]	61.50	62.32	63.12	64.23	65.70
APE-T [22]	**62.51**	**63.25**	**63.66**	64.80	66.07
HC-TR (Ours)	61.46	62.34	62.86	64.47	65.90
HC-TAF (Ours)	61.17	61.79	62.73	64.37	66.04
HC*-TR (Ours)	62.15	62.59	63.60	64.81	66.25
HC*-TAF (Ours)	62.04	62.61	63.28	**64.86**	**66.40**

Table 2: The performance comparison on domain generalization over four CLIP visual backbones. All methods are trained on the ImageNet in 16-shot setting and evaluated on the domain-shifted datasets, ImageNet-V2, -Sketch, -A, and -R.

Method	Backbone	Source	Target				
		ImageNet	-V2	-Sketch	-A	-R	Average
Zero-shot CLIP [1]	ResNet-50	58.18	51.34	33.32	21.65	56.00	40.58
Linear Probe CLIP [1]		55.87	45.97	19.07	12.74	28.16	28.16
TaskRes [20]		65.41	56.84	35.54	21.68	59.96	43.51
Tip-Adapter-F [21]		65.43	57.20	35.99	**23.52**	60.45	44.29
HC-TR (Ours)		65.90	56.97	35.36	21.20	59.57	43.28
HC-TAF (Ours)		**66.04**	**57.44**	**36.17**	23.49	**60.52**	**44.41**
Zero-shot CLIP [1]	ResNet-101	61.62	54.81	38.71	28.05	64.38	46.49
Linear Probe CLIP [1]		59.75	50.05	26.80	19.44	47.19	35.87
TaskRes [20]		68.26	59.94	41.30	28.91	67.36	49.38
Tip-Adapter-F [21]		68.47	59.69	41.63	30.05	68.04	49.85
HC-TR (Ours)		68.62	59.66	41.12	29.07	66.97	49.21
HC-TAF (Ours)		**68.80**	**60.30**	**41.76**	**30.08**	**68.15**	**50.07**
Zero-shot CLIP [1]	ViT-B/32	62.05	54.79	40.82	29.57	65.99	47.79
Linear Probe CLIP [1]		59.58	49.73	28.06	19.67	47.20	36.17
TaskRes [20]		68.45	59.54	42.09	30.60	68.80	50.18
Tip-Adapter-F [21]		68.55	59.10	42.62	32.08	69.53	50.83
HC-TR (Ours)		68.71	59.57	42.09	30.59	68.86	50.28
HC-TAF (Ours)		**69.04**	**59.75**	**42.74**	**32.16**	**69.60**	**51.06**
Zero-shot CLIP [1]	ViT-B/16	66.73	60.83	46.15	47.77	73.96	57.18
Linear Probe CLIP [1]		65.85	56.26	34.77	35.68	58.43	46.29
TaskRes [20]		73.55	65.81	48.86	49.85	77.35	60.47
Tip-Adapter-F [21]		73.77	65.90	49.13	**50.81**	77.96	60.95
HC-TR (Ours)		73.85	65.98	48.88	50.08	77.51	60.61
HC-TAF (Ours)		**74.08**	**66.18**	**49.30**	50.75	**78.01**	**61.06**

Domain Generalization. We investigate the generalization ability of the models constrained by HC under domain shift. We train the models on ImageNet with 16 shots and test the trained models on ImageNet variant datasets ImageNet-V2, ImageNet-Sketch, ImageNet-A and ImageNet-R. As

shown in Tab. 2, HC-TR and HC-TAF outperform their respective baselines on the source dataset ImageNet across four different visual backbones ResNet-50, ResNet-101, ViT-B/32 and ViT-B/16. The performance of HC-TR is slightly inferior to that of TaskRes on ImageNet-Sketch, -A and -R with backbones ResNet-50, ResNet-101 and ViT-B/32, which concurs with TaskRes's reported source-overfitting pitfall. Nonetheless, HC-TR outperforms TaskRes with the more representative ViT-B/16. HC-TAF shows better generalization than Tip-Adapter-F in almost all target domains with various backbones. Experiments demonstrate that the performance improvement is not reliant on the shortcut to overfit the seen data domain.

4.3 Ablation Studies

Constraint terms: track coincidence and deviating perturbation. The homology consistency constraint consists of track coincidence and deviating perturbation, and the effect of homology consistency constraint in efficient transfer comes from their collaboration. Here, we investigate the individual roles of track coincidence and deviating perturbation in tuning. The ablation on ImageNet dataset is shown in Tab. 3, from which we can observe that (1) Without the generalization enhancement in latent spaces brought by DP, the performance of TC alone decreases relative to full homology consistency constraint. (2) Conversely, if we do not guide the coincidence of persistence tracks and only apply DP to track end-point samples, the original performance of baselines will be damaged. This is because DP plays a role of regularization term for TC. Without coinciding persistence tracks by TC, not only the direct constraint on the structural equivalence of latent manifolds is lost, but also the only DP will cause the track end-point samples to randomly deviate from the hetero-modal training samples, which interferes with the downstream tuning and thus performance drop.

Table 3: The ablation studies for the constraint terms, track coincidence (TC) and deviating perturbation (DP).

Baseline	TC	DP	1-shot	2-shot	4-shot	8-shot	16-shot
TaskRes			61.44	62.28	62.78	64.30	65.41
	✓		61.46	62.38	62.95	64.40	65.74
		✓	-	62.14	62.67	64.23	65.33
	✓	✓	-	62.34	62.86	64.47	65.90
Tip-Adapter-F			61.16	61.74	62.65	64.07	65.43
	✓		61.17	61.76	62.70	64.22	65.99
		✓	-	61.76	62.63	63.95	65.17
	✓	✓	-	61.79	62.73	64.37	66.04

Scaling factors of HC gradients in tuning. The hyper-parameters η and ω scale the weight of homology consistency constraint in tuning and λ controls the strength of deviating perturbation within the constraint (Sec. 3.3). We adopt the constant scaling (with η, λ) for TaskRes and adaptive scaling (with ω, λ) for Tip-Adapter-F. In general, performance first increases and then decreases as η (or ω) and λ increase, and the optimal performance is achieved by a specific combination of their values. Taking 16-shot ImageNet as an example, as shown in Tab. 4, the optimal η and λ are 15 (*i.e.*, $0.15/\tau_{\text{CLIP}}$) and 2.5 on HC-TR, the optimal ω and λ are 0.4 and 100 on HC-TAF. In implementation, we first determine the values of scaling factors at 16-shot, and then migrate them to HC*-TR / HC*-TAF and other few-shot settings. The setting of hyper-parameter factors on other datasets follows a similar procedure.

Table 4: The ablation studies for the scaling factors of gradients in homology consistency constraint.

η ($\lambda = 2.5$)	5	10	15	20	25
HC-TR	65.67	65.83	65.90	65.79	65.72
λ ($\eta = 15$)	1.5	2.0	2.5	3.0	3.5
HC-TR	65.83	65.88	65.90	65.87	65.82
ω ($\lambda = 100$)	0.1	0.2	0.3	0.4	0.5
HC-TAF	65.59	65.82	65.96	66.04	65.99
λ ($\omega = 0.4$)	50	100	150	200	250
HC-TAF	66.03	66.04	66.00	66.00	65.98

5 Conclusions, Limitations and Future Work

In this work, we study the generalizability of image-text alignment adjusting in the efficient transfer of VLMs under a low-data regime. We propose to explicitly constrain the structural equivalence of image and text latent manifolds in downstream tuning and design a theoretically well-founded homology consistency constraint based on persistent homology for VLMs transfer. Our method constraint coincides the persistences of homology classes of topological features between image and text manifolds and applies a deviating perturbation to generalize the persistence coincidence to unseen data. Extensive experiments demonstrate the effectiveness and robustness of our method.

Limitations and Future Work. In structural equivalence constraint, we do not explore the effects of higher-dimensional homology classes in depth. Besides, following previous work on efficient transfer learning for VLMs, we only apply the proposed homology consistency constraint on a series of few-shot recognition tasks. As a next step, we will further extend the application scenarios of our method to other VLMs downstream tasks.

Acknowledgments and Disclosure of Funding

This research is supported by Artificial Intelligence-National Science and Technology Major Project 2023ZD0121200, National Science Fund for Excellent Young Scholars No. 62222212 and National Natural Science Foundation of China No. 62336001.

References

[1] 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.

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[3] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004, 2021.

[4] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8552–8562, 2022.

[5] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[6] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9611–9620, 2022.

[7] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022.

[8] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022.

[9] 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, 2022.

[10] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[11] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[12] 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*, pages 19113–19122, 2023.

[13] Sheng Shen, Shijia Yang, Tianjun Zhang, Bohan Zhai, Joseph E. Gonzalez, Kurt Keutzer, and Trevor Darrell. Multitask vision-language prompt tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5656–5667, 2024.

[14] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023.

[15] Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15670–15680, 2023.

[16] Eulrang Cho, Jooyeon Kim, and Hyunwoo J Kim. Distribution-aware prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22004–22013, 2023.

[17] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022.

[18] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. In *International Conference on Learning Representations*, 2023.

[19] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023.

[20] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023.

[21] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pages 493–510. Springer, 2022.

[22] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2605–2615, 2023.

[23] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15211–15222, 2023.

[24] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

[25] James R Munkres. *Elements of algebraic topology*. CRC press, 2018.

[26] Tamal Krishna Dey and Yusu Wang. *Computational topology for data analysis*. Cambridge University Press, 2022.

[27] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36, 2023.

[28] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[29] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831, 2021.

[30] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[31] Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15159–15168, 2023.

[32] Huatian Zhang, Lei Zhang, Kun Zhang, and Zhendong Mao. Identification of necessary semantic undertakers in the causal view for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7105–7114, 2024.

[33] Zheren Fu, Lei Zhang, Hou Xia, and Zhendong Mao. Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26307–26316, 2024.

[34] Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. *Advances in neural information processing systems*, 32, 2019.

[35] James R Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A Zimmer, Julia A Schnabel, and Andrew P King. A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8766–8778, 2020.

[36] Xiaoling Hu, Dimitris Samaras, and Chao Chen. Learning probabilistic topological representations using discrete morse theory. In *The Eleventh International Conference on Learning Representations*, 2022.

[37] Nico Stucki, Johannes C Paetzold, Suprosanna Shit, Bjoern Menze, and Ulrich Bauer. Topologically faithful image segmentation via induced matching of persistence barcodes. In *International Conference on Machine Learning*, pages 32698–32727. PMLR, 2023.

[38] Zuoyu Yan, Tengfei Ma, Liangcai Gao, Zhi Tang, and Chao Chen. Link prediction with persistent homology: An interactive view. In *International conference on machine learning*, pages 11659–11669. PMLR, 2021.

[39] Johanna Immonen, Amauri Souza, and Vikas Garg. Going beyond persistent homology using persistent homology. *Advances in Neural Information Processing Systems*, 36, 2023.

[40] Joshua Southern, Jeremy Wayland, Michael Bronstein, and Bastian Rieck. Curvature filtrations for graph generative model evaluation. *Advances in Neural Information Processing Systems*, 36, 2023.

[41] Jacob Townsend, Cassie Putman Micucci, John H Hymel, Vasileios Maroulas, and Konstantinos D Vogiatzis. Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nature communications*, 11(1):3230, 2020.

[42] Yuankai Luo, Lei Shi, and Veronika Thost. Improving self-supervised molecular representation learning using persistent homology. *Advances in Neural Information Processing Systems*, 36, 2023.

[43] Vincent Peter Grande and Michael T Schaub. Topological point cloud clustering. In *International Conference on Machine Learning*, pages 11683–11697. PMLR, 2023.

[44] Naoki Nishikawa, Yuichi Ike, and Kenji Yamanishi. Adaptive topological feature via persistent homology: Filtration learning for point clouds. *Advances in Neural Information Processing Systems*, 36, 2023.

[45] Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *International conference on machine learning*, pages 7045–7054. PMLR, 2020.

[46] Ilya Trofimov, Daniil Cherniavskii, Eduard Tulchinskii, Nikita Balabin, Evgeny Burnaev, and Serguei Barannikov. Learning topology-preserving data representations. In *The Eleventh International Conference on Learning Representations*, 2022.

[47] Serguei Barannikov, Ilya Trofimov, Nikita Balabin, and Evgeny Burnaev. Representation topology divergence: A method for comparing neural network representations. In *International Conference on Machine Learning*, pages 1607–1626. PMLR, 2022.

[48] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[49] 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.

[50] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[51] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[52] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

[53] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014.

[54] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[55] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[56] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[57] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[58] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[59] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[60] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

[61] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021.

[62] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.

A Simplicial Complex, Homology Class, Persistent Homology and Homotopy Equivalent Assumption

A.1 Simplicial Complex

A simplicial complex consists of a set of simplices, such as points, line segments, triangles, and p-dimensional analogues. A simplex σ with dimension p is the convex hull of a set of $p+1$ affinely independent points $(x_0, \ldots, x_p)$. For $0 \leq p' \leq p$, a p'-face of σ is a p'-simplex that is the convex hull of a non-empty subset. A simplicial complex K consists of a set of finite simplices that satisfy: K contains every face of each simplex in K and for any two simplices $\sigma, \tau \in K$, their intersection $\sigma \cap \tau$ is either empty or a face of both σ and τ. The dimension of K is the maximum dimension of any simplex in K. Given a finite data point set X in metric space (M, d) and a real $a > 0$, the Vietoris-Rips complex $K_a^{\mathrm{Rips}}(X)$ is the set of simplices σ such that $d(x, x') \leq a$ for every pair of vertices of σ, and the Čech complex $K_a^{\mathrm{Cech}}(X)$ is the set of simplices σ such that the closed balls centering its vertices have a non-empty interaction.

A.2 Homology Class

In complex K, let the number of p-simplices be m_p, a p-chain c is a formal sum of p-simplices added with some coefficients, *i.e.*, $c = \sum_{i=1}^{m_p} \alpha_i \sigma_i$. In particular, the p-chains with $\mathbb{Z}_2$-additions form a chain group $\mathrm{C}_p(K)$ where the identity is the chain $0 = \sum_{i=1}^{m_p} 0\sigma_i$, and since $c + c = 0$, the inverse of a chain c is itself. When K is clear from the context, $\mathrm{C}_p(K)$ can also be notated as C_p. Define a boundary operator ∂_p on p-simplex σ as a map that sends σ to the $(p-1)$-chain consisting of σ's $(p-1)$-faces referred as σ 's boundary, *i.e.*, $\partial_p \sigma = \sum_{i=0}^{p} (x_0, \ldots, \hat{x}_i, \ldots, x_p)$ where $\hat{x}_i$ indicates that x_i is omitted. The boundary of a vertex is empty. The ∂_p induces a homomorphism $\partial_p : \mathrm{C}_p \to \mathrm{C}_{p-1}$ that produces a $(p-1)$-chain when extended to a p-chain c through $\partial_p c = \sum_{i=1}^{m_p} \alpha_i \left(\partial_p \sigma_i\right)$. Applying ∂_p to the chain groups, a sequence of homomorphism, $0 = \mathrm{C}_{k+1} \xrightarrow{\partial_{k+1}} \mathrm{C}_k \xrightarrow{\partial_k} \mathrm{C}_{k-1} \cdots \mathrm{C}_1 \xrightarrow{\partial_1} \mathrm{C}_0 \xrightarrow{\partial_0} \mathrm{C}_{-1} = 0$, where 0 denotes a trivial group, is obtained. A p-chain c is a p-cycle if c has empty boundary. All p-cycles together form a cycle group Z_p under the $\mathbb{Z}_2$-addition operation inherited from C_p. Z_p is the subgroup of C_p that is mapped to the 0 of C_{p-1} by ∂_p, *i.e.*, Z_p is the kernel of ∂_p. The image of the boundary operator ∂_p is a subgroup of $(p-1)$-chains, called boundary group $\mathrm{B}_{p-1} = \partial_p\left(\mathrm{C}_p\right)$. For any p-simplex σ, every $(p-2)$-faces is involved in exactly two $(p-1)$-faces in $\partial_p \sigma$, thus $\partial_{p-1} \mathrm{B}_{p-1} = 0$ for $p > 0$ and $\mathrm{B}_{p-1} \subseteq \mathrm{Z}_{p-1}$. The homology group H_p is defined to classify the cycles in Z_p by collecting those cycles that differ by a boundary into the same class. This is achieved by $\mathrm{H}_p = \mathrm{Z}_p / \mathrm{B}_p$, taking the quotient of the Z_p with boundary group B_p. By definition, the elements of H_p are cosets of B_p in Z_p, *i.e.*, $\{c + \mathrm{B}_p \mid c \in \mathrm{Z}_p\}$. For a cycle c, $c + \mathrm{B}_p$ in H_p forms its homology class $[c]$. Two cycles c and c' in the same class $[c] = [c']$ are homologous. In fact, $[c] = [c']$ holds if and only if $c \in c' + \mathrm{B}_p$.

A.3 Persistent Homology

For data point set X in metric space (M, d), define a function $f : M \to \mathbb{R}, f\left((x_0, \ldots, x_p)\right) = \max_{i,j \in \{0,\ldots,p\}} f\left((x_i, x_j)\right)$ on simplices. Then, given a sequence of thresholds $a_1 \leq a_2 \leq \ldots, \leq a_n$, the growing sublevel sets $f^{-1}(-\infty, a]$ at these values give rise to a nested sequence of subcomplexes, $K_{a_1} \subseteq K_{a_2} \subseteq \cdots \subseteq K_{a_n}$, called a filtration $\mathcal{F}$. The inclusions in filtration $\mathcal{F}$ induce a sequence of homomorphisms $h_p^{i,j}$, $\mathrm{H}_p\mathcal{F} : 0 = \mathrm{H}_p\left(K_0\right) \to \mathrm{H}_p\left(K_1\right) \to \cdots \to \mathrm{H}_p\left(K_i\right) \xrightarrow{h_p^{i,j}} \cdots \to \mathrm{H}_p\left(K_j\right) \to \cdots \to \mathrm{H}_p\left(K_n\right) = \mathrm{H}_p(K)$, called a homology module. The images of the homomorphisms $h_p^{i,j}$ are persistent homology groups $\mathrm{H}_p^{i,j}$ for the module, $\mathrm{H}_p^{i,j} = \operatorname{im} h_p^{i,j}$, for $0 \leq i \leq j \leq n$. The non trivial elements of persistent homology groups $\mathrm{H}_p^{i,j}$ consist of homology classes that survive from K_i to K_j, *i.e.*, the homology classes that do not quotient out by boundaries, which implies that $\mathrm{H}_p^{i,j} = \mathrm{Z}_p\left(K_i\right) / \left(\mathrm{B}_p\left(K_j\right) \cap \mathrm{Z}_p\left(K_i\right)\right)$. A non-trivial homology class $\epsilon \in \mathrm{H}_p\left(K_a\right)$ is born at K_i, if $\epsilon \in \mathrm{H}_p^{i,a}$ but $\epsilon \notin \mathrm{H}_p^{i-1,a}$. Likewise, the ϵ dies entering K_j, if $\epsilon \in \mathrm{H}_p^{a,j-1}$ but $\epsilon \notin \mathrm{H}_p^{a,j}$. When a class dies, it may be merged with several classes and raises a new birth. The persistence of homology class that is created at K_i and destoryed at K_j is defined as $a_i - a_j$. The homology classes that never die are set to remain alive till $a_{n+1} = a_\infty = \infty$. As a visual representation, the persistence diagram

$\mathrm{Dgm}_p(\mathcal{F})$ of filtration $\mathcal{F}$ draws the paired birth at a_i and death at a_j that bound the survival interval of one or more homology classes as a point (a_i, a_j) on the extended plane $\overline{\mathbb{R}}^2 := (\mathbb{R} \cup \{\pm\infty\})^2$.

A.4 Assumption on Homotopy Equivalent of Image and Text Latent Manifolds

Topology specifies how points are connected and is a key geometric signature of metric spaces. Topological equivalence can be formalized by continuous functions that map points from one space to the other while preserving the connectivity. A straightforward equivalence is that the points as well as their neighborhoods in two spaces are in one-to-one correspondence, called being homeomorphic [24, 26], which requires a bijecitve map with continuous inverse. Two homeomorphic spaces share the exact same topological properties, *i.e.*, from the topological point of view, the two are completely consistent. Another less rigid relation is homotopy equivalent [24, 26]. Formally, given spaces $\mathbb{X}$ and $\mathbb{Y}$, two maps $f_0, f_1 : \mathbb{X} \to \mathbb{Y}$ are homotopic if they can be joined by a continuous function $F : \mathbb{X} \times [0, 1] \to \mathbb{Y}$ such that $F(x, 0) = f_0(x)$ and $F(x, 1) = f_1(x)$ for every $x \in \mathbb{X}$, denoted as $f_0 \simeq f_1$. The spaces $\mathbb{X}$ and $\mathbb{Y}$ are homotopy equivalent if there exist two maps $f : \mathbb{X} \to \mathbb{Y}$ and $g : \mathbb{Y} \to \mathbb{X}$ such that $f \circ g \simeq \mathrm{id}_{\mathbb{Y}}$ and $g \circ f \simeq \mathrm{id}_{\mathbb{X}}$, which suggests an intuitive fact that space $\mathbb{X}$ is homotopy equivalent to $\mathbb{Y}$ if and only if both $\mathbb{X}$ and $\mathbb{Y}$ are deformation retracts of a common space $\mathbb{Z}$.

Vision and language are key information modalities for human cognition and have intrinsic correspondence on semantics. Observe that similar image scenes tend to be described by synonymic texts, that is, the connectivity of image data samples is preserved after mapping them to the text latent manifold, and vice versa. This preservation of connectivity suggests that the two manifolds have similar topological structures. Besides, image and text data are both relevant and complementary concrete expressions of the real world. The latent manifolds in which they reside can be seen as deformation retracts of the general world knowledge space. Thus, it is rational to consider that the image latent manifold and text latent manifold are homotogy equivalent. As a topological invariant of homotogy equivalent spaces, the homology is a viable tool for quantifying and aligning the structure of image and text latent manifolds.

B Full numerical results on few-shot learning

Table 5: Full numerical results of performance comparison on few-shot learning.

Method	Setting	*ImageNet*	*Caltech101*	*DTD*	*EuroSAT*	*FGVCAircraft*	*Flowers102*	*Food101*	*OxfordPets*	*StanfordCars*	*SUN397*	*UCF101*
Zero-Shot CLIP	1-shot	58.18	86.29	42.32	37.56	17.28	66.14	77.31	85.77	55.61	58.52	61.46
CLIP-Adapter		61.20	88.60	45.80	61.40	17.49	73.49	76.82	85.99	55.13	61.30	62.20
TaskRes		61.44	89.49	49.65	63.52	21.36	79.09	75.03	84.60	59.78	61.87	64.26
Tip-Adapter-F		61.16	89.21	49.17	59.10	20.46	79.29	77.52	86.43	59.69	61.95	65.05
HC-TR		61.46	89.78	49.88	63.43	21.39	79.37	75.25	84.66	59.68	62.13	64.16
HC*-TR		62.15	89.61	49.70	64.02	21.48	80.02	75.47	84.57	59.84	62.92	64.79
HC-TAF		61.17	89.57	50.00	61.70	20.67	80.23	77.59	86.64	59.61	62.03	65.69
HC*-TAF		62.04	89.33	50.77	60.65	21.39	80.39	77.89	86.84	59.56	63.75	66.51
Zero-Shot CLIP	2-shot	58.18	86.29	42.32	37.56	17.28	66.14	77.31	85.77	55.61	58.52	61.46
CLIP-Adapter		61.52	89.37	51.48	63.90	20.10	81.61	77.22	86.73	58.74	63.29	67.12
TaskRes		62.28	89.86	54.43	65.99	23.67	83.09	75.88	85.64	62.70	64.89	67.41
Tip-Adapter-F		61.74	89.86	54.37	66.19	23.34	85.30	77.73	87.05	61.52	64.25	68.28
HC-TR		62.34	89.90	54.79	66.91	24.63	83.11	76.03	85.66	63.06	65.04	67.57
HC*-TR		62.59	90.14	55.67	66.22	25.20	85.71	75.81	85.83	64.48	65.41	69.18
HC-TAF		61.79	90.26	54.61	67.53	24.21	85.99	77.78	87.22	63.11	64.42	68.31
HC*-TAF		62.61	90.63	56.09	67.33	24.69	87.49	77.65	88.14	64.48	65.77	69.94
Zero-Shot CLIP	4-shot	58.18	86.29	42.32	37.56	17.28	66.14	77.31	85.77	55.61	58.52	61.46
CLIP-Adapter		61.84	89.98	56.86	73.38	22.59	87.17	77.92	87.46	62.45	65.96	69.05
TaskRes		62.78	90.71	59.40	74.42	26.49	89.97	76.45	87.03	66.60	67.22	69.36
Tip-Adapter-F		62.65	90.95	58.22	73.46	25.98	89.69	78.12	87.49	65.73	67.13	71.05
HC-TR		62.86	90.83	59.40	75.02	26.76	90.86	76.49	87.10	66.98	67.34	69.92
HC*-TR		63.60	91.60	59.34	75.14	27.69	92.13	76.86	87.22	68.54	67.36	71.64
HC-TAF		62.73	91.44	58.87	78.54	28.65	90.99	78.20	87.71	67.07	67.06	72.03
HC*-TAF		63.28	91.76	59.99	78.74	28.80	91.80	78.53	88.17	68.29	67.84	73.17
Zero-Shot CLIP	8-shot	58.18	86.29	42.32	37.56	17.28	66.14	77.31	85.77	55.61	58.52	61.46
CLIP-Adapter		62.68	91.40	61.00	77.93	26.25	91.72	78.04	87.65	67.89	67.50	73.30
TaskRes		64.30	92.09	63.48	78.07	32.25	94.05	77.82	87.35	70.70	69.29	74.46
Tip-Adapter-F		64.07	92.25	63.12	78.04	32.16	91.76	78.52	87.79	70.12	69.43	74.97
HC-TR		64.47	92.41	63.36	78.35	32.34	94.19	77.98	87.79	71.38	69.37	74.81
HC*-TR		64.81	92.62	64.30	78.96	33.57	94.88	77.99	88.12	73.03	69.39	75.84
HC-TAF		64.37	92.45	63.83	81.35	34.62	93.02	78.92	88.61	71.98	69.74	76.02
HC*-TAF		64.86	93.06	65.13	81.83	35.52	94.36	78.86	89.13	73.39	69.88	77.24
Zero-Shot CLIP	16-shot	58.18	86.29	42.32	37.56	17.28	66.14	77.31	85.77	55.61	58.52	61.46
CLIP-Adapter		63.59	92.49	65.96	84.43	32.10	93.9	78.25	87.84	74.01	69.55	76.76
TaskRes		65.41	93.31	67.02	83.25	35.82	95.53	78.33	88.53	77.07	71.41	78.43
Tip-Adapter-F		65.43	93.06	66.55	84.98	36.42	94.64	79.4	89.67	75.80	71.52	78.09
HC-TR		65.90	93.59	67.61	83.19	37.20	96.06	78.57	88.93	77.61	71.64	78.54
HC*-TR		66.25	93.75	67.85	83.78	38.25	95.98	78.49	88.96	77.75	71.34	78.20
HC-TAF		66.04	93.35	68.32	86.65	38.34	95.37	79.54	90.49	77.55	72.01	79.22
HC*-TAF		66.40	93.83	68.56	85.73	38.82	95.82	79.61	90.24	77.64	71.62	79.54

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We clearly express our motivation and insight in the Abstract and Introduction. The contributions are summarized in the Introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We clearly express our limitations in the Sec. 5, Conclusions, Limitations and Future Work.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

Answer: [Yes]

Justification: We discuss the homotopy equivalent assumption underpinning our method in Appendix A.4 and explain the rationale for such an assumption.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: Our paper provides a detailed account of implementation details and hyper-parameter settings of our method for reproducing.

   Guidelines:

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

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

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

Answer: [No]

Justification: We will make all code avaliable when paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

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

Answer: [Yes]

Justification: We provide implementation details in Sec. 4.1, Experimental Settings.

Guidelines:

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

7. **Experiment Statistical Significance**

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

Answer: [No]

Justification: We fix the random seeds in experiments for fair comparison.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: We specify the type of compute workers in implementation details in Sec. 4.1.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [NA]

    Justification: We believe that our study will not pose any negative societal due to its theoretical nature.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: Our paper does not contain data or models that are at high risk of misuse.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite the datasets used in our paper and give a brief introduction.

    Guidelines:

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

13. **New Assets**

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

    Answer: [NA]

    Justification: We does not release new assets.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: Our paper does not involve human subjects.

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: Our paper has nothing to do with crowdsourcing and human subjects.

    Guidelines:

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