FreestyleRet: Retrieving Images from Style-Diversified Queries

Hao Li[1,2*], Yanhao Jia[3,4*], Peng Jin[1,2], Zesen Cheng[1], Kehan Li[1], Jialu Sui[5], Chang Liu[6], and Li Yuan[1,2†]

[1] School of Electronic and Computer Engineering, Peking University
[2] Peng Cheng Laboratory, Shenzhen, China
[3] College of Computing and Data Science, Nanyang Technological University
[4] Deep NeuroCognition Lab, I2R and CFAR, Agency for Science, Technology and Research, Singapore
[5] School of Science and Engineering, Chinese University of Hong Kong
[6] Department of Automation, Tsinghua University
{lihao1984, yuanli-ece}@pku.edu.cn, {jiayanhao.publish, jp21mails}@gmail.com, {cyanlaser, kehanli}@stu.pku.edu.cn, jialusui@link.cuhk.edu.cn, liuchang2022@tsinghua.edu.cn

Abstract. Image Retrieval aims to retrieve corresponding images based on a given query. In application scenarios, users intend to express their retrieval intent through various query styles. However, current retrieval tasks predominantly focus on text-query retrieval exploration, leading to limited retrieval query options and potential ambiguity or bias in user intention. In this paper, we propose the Style-Diversified Query-Based Image Retrieval task, which enables retrieval based on various query styles. To facilitate the novel setting, we propose the first Diverse-Style Retrieval dataset, encompassing diverse query styles including text, sketch, low-resolution, and art. We also propose a light-weighted style-diversified retrieval framework. For various query style inputs, we apply the Gram Matrix to extract the query's textural features and cluster them into a style space with style-specific bases. Then we employ the style-init prompt learning module to enable the visual encoder to comprehend the texture and style information of the query. Experiments demonstrate that our model outperforms existing retrieval models on the style-diversified retrieval task. Moreover, style-diversified queries (sketch+text, art+text, etc) can be simultaneously retrieved in our model. The auxiliary information from other queries enhances the performance within the respective query, which may hold potential significance for the community. [1]

Keywords: Style-Diversified Image Retrieval · Prompt Learning

1 Introduction

Query-based image retrieval (QBIR) [36] refers to the task of retrieving relevant images from a large image database based on the user's query or search term.

[1] † corresponding author. ∗ equal contribution. Code and Dataset available in here.

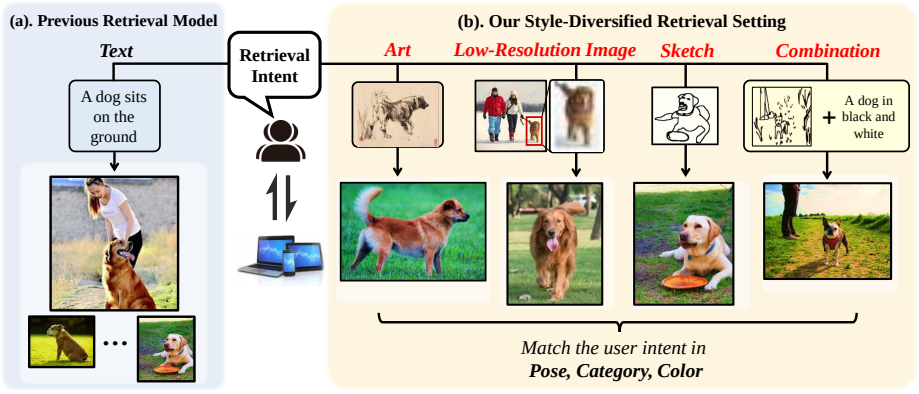

Fig. 1: (a). Previous Retrieval Models focus on text-query retrieval exploration, neglecting the retrieval ability for other query styles. (b). Our style-diversified retrieval setting considers the various query styles that users may prefer, including sketch, art, low-resolution, text, and their combination. Our model makes fine-grained retrieval based on the shape, color, and pose features from style-diversified query inputs.

QBIR has numerous applications, ranging from image search engines [13] to cross-modality downstream tasks [21,22]. It plays a crucial role in enabling users to locate and obtain related visual content based on their retrieval intent.

The diversification of user retrieval intents poses a significant and unresolved problem in QBIR [26]. Selecting appropriate queries to express user intents and enabling models to accommodate diverse query styles are crucial challenges. However, the current exploration in the field of QBIR has primarily focused on text-image retrieval [24,32] and text-video retrieval [15,16], with less emphasis on other query types [17]. To address the issue of limited query style adaptability in current retrieval models, we propose a novel setting: style-diversified query-based image retrieval in Fig. 1(b). The objective of this setting is to enable retrieval models to simultaneously accommodate various query styles, aiming to bridge the user intent gap caused by the lack of query adaptation versatility.

We propose the diverse-style retrieval dataset (DSR) as the evaluation dataset of our style-diversified QBIR task. As shown in Fig. 2(a), the dataset contains 10,000 natural images and four corresponding query styles: text, sketch, low-resolution, and art. *(i). Text:* the text-form query to describe the retrieval intent. *(ii). Sketch:* hand-drawn sketch by users to provide shape and pose features. *(iii). Low-Res:* users capture regions of interest from images and convert them into low-resolution images to serve as queries. *(iv). Art:* artistic-style images as retrieval queries.

We further propose a lightweight plug-and-play framework, FreestyleRet, for the style-diversified retrieval task. For query inputs with different styles, we calculate each query's Gram Matrix [2,27] as the query's style representation, due to the Gram Matrix's ability to capture the textural information and spa-

tial relationships between channels in the input image. Then, we construct the high-dimensional style space by clustering all-style queries' gram matrices and taking the clustering centers as the style basis in the style space. With the well-constructed style space, we introduce the application of a style-init prompt tuning module on a frozen visual encoder [24, 32], thereby enabling the encoder to adapt to various-style queries in a cost-effective manner. Specifically, given a query input, we employ its corresponding Gram matrix in conjunction with the weighted projections within the style space onto the diverse style basis as the initialization mechanism for prompt tokens in the prompt tuning procedure. Finally, we use the query feature from the visual encoder for further retrieval.

The proposed framework has three compelling advantages: **First,** The style-space construction and the style-init prompt tuning strategy enable the framework to adapt to various query styles. Experimental results on two benchmark datasets demonstrate the advantages of our model. **Second,** Our framework is compatible with the retrieval of multiple query types simultaneously, thereby promoting the single-query retrieval performance. **Third,** the prompt tuning structure lowers the computation cost and achieves plug-and-play abilities on various pre-trained visual encoders. The main contributions are as follows:

- We are the first to propose the style-diversified QBIR task and the dataset, DSR, to address the users' intent gap problem in retrieval applications.
- Our framework is lightweight and plug-and-play. With the style space construction module and the style-init prompt tuning module, our framework achieves excellent performance when retrieving style-diversified queries.
- More encouragingly, the style-diversified queries can be simultaneously retrieved in our framework and mutually enhance each other's performance, which may have a far-reaching impact on the retrieval community.

2 Related Works

Query-based Image Retrieval. Query-based Image Retrieval (QBIR) [36] aims to retrieve relevant images from a large database based on a given query. In QBIR, the query can take different forms. The earliest query form is images including natural-image retrieval [7] and face retrieval [18]. With the development of cross-modal representation learning, text-style query tasks are extensively investigated, including text-image retrieval [24,32] and text-video retrieval [15,16]. Limited research incorporates other query styles such as sketch [5, 6] and scene graph [17]. Compared to conventional tasks that retrieve from one given query, our style-diversified retrieval setting allows retrieving from text, sketch, art, and low-resolution queries simultaneously. Thus, establishing a pipeline that can understand style features and extract semantic features from style-diversified queries is the main challenge of our setting.

Prompt Tuning. The objective of Prompt Tuning [20, 25] is to enhance the transferability of pre-trained models to downstream tasks in a cost-effective manner by incorporating learnable tokens into the fixed pre-trained models. Prompt Tuning was first proposed as text-prompt [3, 29] in the language model and then

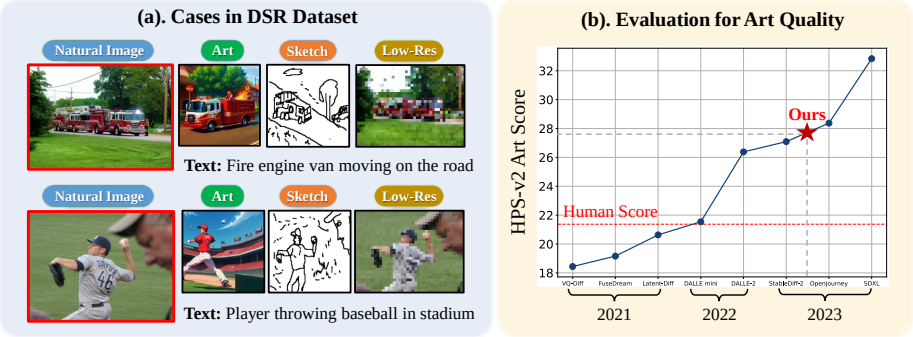

Fig. 2: We propose the **Diverse-Style Retrieval Dataset (DSR)**, containing 10,000 natural images and their corresponding queries with various styles, including Sketch, Art, Low-Resolution (Low-Res), and Text. We make art quality evaluation by HPS-v2 metric [40] to demonstrate the quality of our generated art queries.

gained popularity [14] in vision-language models. Specifically, CoOP [46] and CoCoOP [45] learn class-specific continuous prompts. MaPLe [19] further transfers text features to the visual encoder during prompt tuning to avoid overfitting. However, these methods focus on class-level transferability, having limitations on extracting semantic features from style-diversified images.

Multi-Task Learning and Domain Adaptation. Multi-task learning [43] is a learning paradigm aiming to learn multiple tasks simultaneously. Compared to other similar paradigms including domain adaptation [10] and domain generalization [44], the data of each task in multi-task learning is well-labeled. In our style-diversified QBIR setting, we collect abundant well-labeled images for each image style, which can enable the fine-tuning and convergence of CLIP-level models. Thus, our style-diversified QBIR setting belongs to the multi-task learning paradigm. From the methodology aspect, prompt learning methods, including CoCoOP [45], CoOP [46], MaPLe [19], are commonly used in both domain adaptation and multi-task learning [30, 41]. Thus, We apply them as baselines for a fair comparison. A detailed comparison between the multi-task learning (our setting) and the domain adaptation task is provided in the supplementary material.

3 Dataset Construction

In the context of the style-diversified QBIR task, we adopt two datasets for evaluation: the Diverse-Style Retrieval dataset and ImageNet-X.

Diverse-Style Retrieval Dataset: A small but fine-grained dataset constructed for style-diversified QBIR. Shown in Fig. 2(a), it consists of 10,000 natural images paired with corresponding queries of four styles: text, sketch, low-res, and art. *(i). Text:* the text query used to express the retrieval intent. *(ii). Sketch:* hand-drawn sketch by users to provide shape and pose features.

(iii). Low-Res: users capture regions of interest from images and convert them into low-resolution images to serve as queries. *(iv). Art:* artistic-style images as queries. With the rise of the AIGC [4,39,42], generating images of different styles has become more convenient. Therefore, based on ten thousand natural images from FSCOCO [6], we utilize AnimateDiff [11] to generate corresponding artistic style images. We employ downsampling algorithms to generate low-resolution images. As for sketch images, FSCOCO provides high-quality sketch images for each natural image.

To demonstrate the validity of our Diversified-Style Retrieval dataset, we make a user study for the style-diversified queries. Both sketch and low-resolution queries are close to the user input because the sketch images in FSCOCO are drawn by 100 non-expert participants and the resolution of our low-resolution queries is similar to the user's image segmentation. As for the art queries in the DSR dataset, we apply a solid image quality benchmark, HPS-V2 [40], to score our art queries. Based on Fig. 2(b), we can demonstrate that the artistic queries in the DSR dataset exhibit a higher aesthetic quality compared to the art drawn by humans.

ImageNet-X: A large but coarse-grained dataset for style-diversified QBIR. Based on ImageNet [8], ImageNet-X contains 1M natural images and their corresponding sketch-form and art-form versions. Compared to DSR, the images in ImageNet-X are simple, containing only one object. We generate the low-resolution form for images and reconstruct ImageNet-X as the dataset for style-diversified QBIR.

4 Methodology

Given a gallery of natural images N_I and a query q_i from the style-specific query set Q_s. The goal for query-based image retrieval is to rank all images $i \in N_I$ so that the image corresponding to the query q_i is ranked as high as possible. For our style-diversified QBIR setting, the goal is similar, ranking all images correctly with queries for various style-specific query sets $\{Q_s\}_{s=1}^{n}$.

Our model consists of three main submodules: (1) a **Gram-based Style Extraction Module** for generating the gram matrix of an input query, representing the query's textural feature (Sec.4.1). (2) a **Style Space Construction Module** for building up the query style space by clustering queries' gram matrices and taking the cluster centers as the style basis (Sec.4.2). (3) a **Style-Init Prompt Tuning Module** for style-specific prompt tuning a pre-trained visual encoder by initializing the prompt tokens based on the gram matrices and the style prototypes (Sec.4.3). The overview framework of our FreestyleRet is illustrated in Fig. 3.

4.1 Gram-based Style Extraction Module

For query inputs with diverse styles, it's vital to distinguish and extract their different style features. Considering the outstanding representation of the image style provided by the Gram matrix [2, 27], we propose the gram-based style

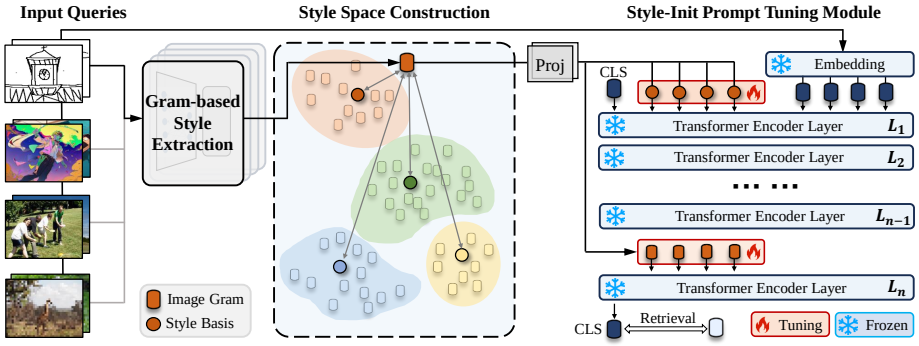

Fig. 3: The Overall Framework of our FreestyleRet. For a style-diversified query input, we first extract the query's textural feature by calculating the query's gram matrix from the Gram-based Style Extraction Module. Then we construct the style space of queries by clustering all gram matrices and taking each clustering center as the style basis in style space. We further extract the query's style feature by weighted summarizing style bases based on the distance between the input query and every style basis in the style space. Finally, in the Style-Init Prompt Tuning Module, we use the gram matrix and the style feature to initialize prompt tokens, leading both textural and style information to the feature encoder for further style-diversified retrieval prediction.

extraction module to capture the style representation for input queries by calculating their Gram matrix.

First, we apply the frozen VGG model [33] to get the query's visual feature. Compared with other image feature extractors including ViT [9] and ResNet [12], the VGG model is lightweight and has strong feature extraction ability during the gram matrix calculation in image style transfer works [35, 38]. The VGG model is constituted by a concatenation of 16 layers, consisting of stacked convolutional and fully connected layers, meticulously structured to capture complex patterns in the visual data. For the query input q_i, we use the third convolutional layer output, shaping $112 \times 112 \times 128$, as the visual feature v_i of the query q_i. $f_d(.)$ is used to downsample v_i.

Then, we calculate the gram matrix for query q_i. Specifically, the Gram Matrix g of a set of vectors $t_1, ..., t_n$ in an inner product space is the Hermitian matrix of inner products: $g_{jk} =< t_j, t_k >$. g represents the texture feature of vectors $t_1, ..., t_n$. In our scenario, we calculate the gram matrix g_i for q_i as follows:

$$g_i = (f_d(v_i))^{\mathsf{T}} f_d(v_i), \tag{1}$$

where g_i represents the textural feature of the query q_i.

4.2 Style Space Construction Module

For style-diversified query inputs, we construct the style space $\mathbb{S}$ for queries to encode their specific styles. To generate the style-specific basis $\mathbb{B} = \{b_j\}_{j=1}^{4}$ for

the style space, we cluster the gram matrices of all queries in various styles and apply each clustering center as the style-specific basis b_j for the style space $\mathbb{B}$.

During the clustering procedure, we apply the K-Means algorithm to cluster the gram matrix set G for all queries from query sets in the dataset, where $G = \{g_i\}, \forall q_i \in Q_s$. We first random initialize four clustering centers $\mu_1, ..., \mu 4$ as the basis of the style space. Then we calculate the nearest center c_i comparing each gram matrix $g_i \in G$ with existing clustering centers:

$$c_i = \arg\max_j ||g_i - \mu_j||^2, \tag{2}$$

where $j = 1, ..., 4$. We redistribute all queries to their nearest center based on the c_i. Then we refine the position of μ_j by averaging all queries belong to μ_j:

$$\mu_j = \frac{\sum_{i=1}^{m} \mathbf{Num}\{c_i = j\} \times g_i}{\sum_{i=1}^{m} \mathbf{Num}\{c_i = j\}}, \tag{3}$$

We repeat the iteration of Eq.2 and Eq.3 until the clustering centers' positions converge. The well-trained clustering centers $\mu_1, ..., \mu_4$ act as the style-specific basis for the constructed style space. We further use these style-specific bases to represent the style feature s_i of an input query q_i. Specifically, the style feature s_i is calculated by weighted summarizing all the style bases according to the cosine similarity w between q_i and $\mu_j, \forall j \in [1, 4]$.

$$w_j = \frac{e^{\cos(q_i, \mu_j)}}{\sum_{j=1}^{4} e^{\cos(q_i, \mu_j)}}, \quad s_i = \sum_{j=1}^{4} w_j \mu_j, \tag{4}$$

The weighted summarizing calculation enables the model to generate the q_i's style feature adaptively.

4.3 Style-Init Prompt Tuning Module

To build up a lightweight and plug-and-play framework, we apply the prompt tuning procedure on a frozen pre-trained visual encoder to make the frozen visual encoder understand the various-style query inputs. As shown in Fig. 3, during the prompt tuning, we insert four trainable prompt tokens into both the shallow layer and the bottom layer of the vision transformer encoder, to tune the visual encoder comprehensively. The prompt tokens are introduced to every transformer layer's input space. For i-th Layer L_i in the transformer, we denote the collection of input learnable prompts P_i as

$$P_i = \{p_i^k \in \mathbb{R}^d | k \in \mathbb{N}, 1 \leq k \leq m\}, \tag{5}$$

where $d = 1024$ represents the token dimension in the transformer layer. $m = 4$ represents the prompt token number for each transformer layer. The style-init prompt tuning module for ViT is formulated as follows:

$$[x_i, \_, E_i] = L_i(x_{i-1}, P_{i-1}, E_{i-1}), i = 1, ..., n \tag{6}$$

$$f_i = \mathbf{Head}(x_n), \tag{7}$$

where n represents the transformer layer number, x_i represents the [CLS]'s embedding at L_i's input space, E_i is q_i's image patch embeddings. **Head** represents the MLP to generate visual feature f_i using the [CLS] embedding of q_i.

We apply style-specific initialization for prompt tokens in both shallow and deep layers in the visual encoder to achieve multi-scale style feature fusion. Specifically, given an input query q_i, we initialize the prompt tokens in the shallow layer based on the gram matrix from Eq.1 and initialize the tokens in the deep layer based on the style feature s_i calculated from Eq.4. Further experimental analysis in Table. 3 shows that differentiated style initialization across different layers can boost the performance of the ViT-based visual encoder.

4.4 Training and Inference

As shown in Fig. 3, our FreestyleRet iterates the dataset twice during the training process. We first construct the style space during the first iteration. Then we apply the well-constructed style space for style-init prompt tuning during the second iteration. The overall loss $\mathcal{L}$ of our model is the triplet loss:

$$\mathbf{dist}(x, y) = 1 - \cos(x, y), \tag{8}$$

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B}(\max(0, \mathbf{dist}(F_i, P_i) - \mathbf{dist}(F_i, N_i) + \alpha)) \tag{9}$$

where F represent the image features $F = \{f_i\}_1^n$. P represents the positive samples and N represents the negative samples. During the training, we take the ground-truth retrieval answer as P. For N we randomly select another image from the same query-style set as q_i. We set the hyperparameter α to 1.0.

Our inference process iterates the test dataset once, using the gram-based style extraction module and the well-constructed style space to get the textural feature from the gram matrix and style feature for the input query. Then we apply the style-init prompt tuning module for retrieval.

5 Experiments

5.1 Experimental Baselines and Settings

For the baseline selection, we apply two multi-modality pre-trained models (ImageBind, LanguageBind), two cross-modality pre-trained models (CLIP, BLIP), and the three most recent cross-modality prompt learning models (VPT, CoCoOP, MaPLe) for the fair comparison. Specifically, we fine-tune the cross-modality models (CLIP, BLIP) on DSR and ImageNet-X datasets for convergence. we also train the prompt learning baselines on both datasets to adapt style-diversified inputs. As for the multi-modality models, we evaluate the zero-shot performance on the sty-diversified retrieval task due to multi-modality models' comprehensionability on multi-style image inputs.

Table 1: Retrieval performance for Style-Diversified QBIR task. We evaluate the R@1 and R@5 metrics on two benchmark datasets, the Diverse-Style Retrieval dataset and the ImageNet-X dataset. The two forms of our FreestyleRet framework, FreestyleRet-CLIP and FreestyleRet-BLIP, outperform in multiple scenarios with different query styles compared with other baselines including multi-modality models, cross-modality pre-trained models, and prompt learning models.

#	Method	Text → Image		Sketch → Image		Art → Image		Low-Res → Image	
		R@1↑	R@5↑	R@1↑	R@5↑	R@1↑	R@5↑	R@1↑	R@5↑
Diverse-Style Retrieval Dataset									
1	CLIP-ZeroShot	66.1	94.7	47.5	77.3	58.5	93.7	45.0	75.7
2	CLIP-Finetune	73.8	97.4	78.4	95.6	70.2	96.6	84.0	97.1
3	BLIP-Finetune	80.1	98.7	75.3	94.8	63.7	93.4	83.7	96.3
4	ImageBind	71.0	95.5	50.8	79.4	58.2	86.3	79.0	96.7
5	LanguageBind	79.7	98.1	63.6	89.1	67.5	92.9	78.6	94.5
6	VPT	69.9	96.1	73.3	97.0	66.7	96.5	81.4	96.0
7	CoCoOP	71.4	94.6	77.5	97.2	69.3	97.1	83.8	97.6
8	MaPLe	73.1	95.9	80.3	97.9	70.6	97.2	85.9	97.7
9	**FreestyleRet(C)**	71.4	97.2	81.6	98.0	72.3	**98.1**	86.7	98.2
10	**FreestyleRet(B)**	**81.6**	**99.2**	**82.0**	**98.4**	**74.5**	97.4	**90.5**	**98.5**
ImageNet-X Dataset [8]									
11	CLIP-Finetune	42.6	72.7	41.3	73.9	38.5	65.3	74.1	95.7
12	BLIP-Finetune	63.9	90.7	53.6	88.1	49.6	84.8	89.5	97.8
13	ImageBind	57.3	89.7	53.6	86.2	49.8	79.3	81.2	94.3
14	LanguageBind	68.9	92.3	62.0	91.5	60.3	89.9	87.4	99.5
15	VPT	43.3	85.3	48.6	84.2	41.6	88.5	72.7	89.3
16	CoCoOP	64.4	91.7	54.8	90.4	52.6	86.6	73.9	95.0
17	MaPLe	65.2	94.8	56.2	87.5	53.4	89.3	74.2	96.2
18	**FreestyltRet(C)**	64.8	94.3	57.7	90.5	56.4	90.2	77.4	96.5
19	**FreestyleRet(B)**	**74.9**	**96.3**	**74.6**	**93.3**	**71.2**	**96.5**	**97.5**	**99.7**

For the experiments on the DSR and the ImageNet-X datasets, FreestyleRet is trained on one A100 GPU with batch size 24 and 20 training epochs. The learning rate is set to 1e-5 and is linearly warmed up in the first epochs and then decayed by the cosine learning rate schedule. All input images are resized into 224×224 resolution and then augmented by normalized operation.

5.2 Main Results

We apply two benchmark datasets, the ImageNet-X and the DSR dataset, for our style-diversified retrieval task. The results in Tab. 1 yield three observations:

(i). Cross-modality pre-trained models, prompt models, and multi-modality models have the potential for improvement in the style diversified retrieval task. Line.1 in Tab. 1 shows that zero-shot CLIP performs badly compared with our FreestyleRet. This limitation arises from the inability of vision-linguistic models like CLIP to distinguish visual inputs with different styles from those of natural images in the feature space. With the finetuning process, pre-trained models have significant improvements, as shown in line.2-3 and line.11-12. As for the multi-modality models, ImageBind and LanguageBind,

Table 2: Retrieval performance with multi-style queries simultaneously. The additional query inputs (sketch, art, low-res) can boost the text-image retrieval capability in our FreestyleRet while showing a negative influence on baseline models, including CLIP and BLIP.

#	Method	Sketch		Art		Low-Resolution	
		Text→I	Sketch+Text→I	Text→I	Art+Text→I	Text→I	Low-Res+Text→I
1	CLIP-finetune	72.2	65.0$_{(-7.2)}$	72.2	57.8$_{(-14.4)}$	72.2	84.7$_{(+12.5)}$
2	BLIP-finetune	74.3	74.2$_{(-0.1)}$	74.3	58.3$_{(-16.0)}$	74.3	**88.3**$_{(+14.0)}$
3	**FreestyleRet**	71.4	**82.5**$_{(+12.6)}$	71.4	**76.6**$_{(+6.7)}$	71.4	**86.7**$_{(+16.8)}$

line.4-5 and line.13-14 show that multi-modality models have style-diversified retrieval abilities. Shown in line.6-8 and 15-17, the prompt learning models also have style-diversified retrieval abilities but still need improvements when facing sketch and art query styles.

(ii). The CLIP-form and BLIP-form models of our FreestyleRet framework outperform both cross-modality and multi-modality models. Claimed in Sec.4.3, our FreestyleRet is a plug-and-play framework that can easily applied to various pretrained visual encoders. Here we apply our FreestyleRet on two ViT-based visual encoders from CLIP and BLIP. We use FreestyleRet-CLIP and FreestyleRet-BLIP as the generated models. Line.7-8 and line.14-15 in Tab. 1 show that both FreestyleRet-CLIP and FreestyleRet-BLIP outperform the cross-modality and multi-modality baselines, demonstrating the effectiveness of our plug-and-play framework.

(iii). FreestyleRet-BLIP outperforms its CLIP-form by a large margin. Shown in line.7-8 and 14-15, FreestyleRet-BLIP performs better than its CLIP form. Proved in [23, 24], BLIP applies the MED structure and cleaner datasets for pretraining, which leads to a better generalization ability than CLIP on various modalities. Thus, with the extracted style feature as prompt initialization, BLIP can be generalized to style-diversified image groups more easily, leading to better performance.

(iv). In our FreestyleRet framework, style-diversified queries can be simultaneously retrieved and mutually enhance the text-image retrieval performance. As shown in Tab. 2, when conducting text-image retrieval, the additional query inputs (sketch, art, low-res) can significantly boost the text-image retrieval capability of our FreestyleRet framework. However, for baseline models, the additional query signals cannot stably improve the text-image retrieval performance. In line.1-2 in Tab. 2 the additional sketch and art queries hurt the CLIP and BLIP.

5.3 Ablation Studies

In the ablation section, we first make an adequate ablation analysis for the prompt tuning structure to validate the rationality of our model design. We

also compare our FreestyleRet with baselines from the aspect of computational complexity to demonstrate the lightweight nature of our model. Moreover, we also verified the state-of-the-art performance of our model in the standalone task of sketch retrieval.

Ablation for Prompt Tuning Structure. We ablate the prompt tuning structure in our FreestyleRet framework from two aspects: the prompt token initialization feature and the prompt token number. Table. 3 shows the ablation results. Furthermore, Fig. 4 proposes the detailed structure of the prompt tuning module in FreestyleRet.

The prompt token position. Previous prompt tuning models [14, 28, 29, 31] analyzed that inserting the learnable prompt tokens in all layers in the transformer (Deep Prompt) has better performance than in the first layer in the transformer (Shallow Prompt). In the prompt module of our FreestyleRet, we adopt the deep prompt idea and insert all the learnable prompt tokens into all layers.

The prompt token initialization. We analyze the impact of the prompt token initialization by applying different initialization strategies in different positions of the visual encoder. Line.1-5 in Table. 3 show the ablation results, where "Random" represents random initialization, "Gram" represents initializing with textual information from the gram matrix, and "Style Space" represents initializing with style information from the style space feature. The random initialization in line.1 performs worst, demonstrating that applying textural and style representation as initialization is necessary. We make various initialization attempts in line.2-4 and find that initializing the shallow-layer prompt tokens with style features, while initializing the deep-layer prompt tokens with gram matrices, achieves the best performance.

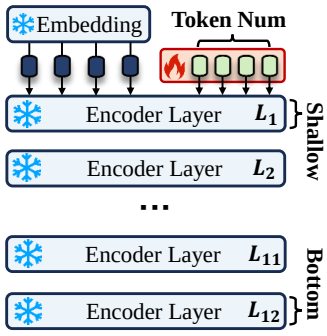

Fig. 4: Prompt tuning structure in FreestyleRet.

The prompt token number. We make ablation studies for the number of prompt tokens that are inserted into the visual encoder during the prompt tuning stage. As shown in line.5-8 in Table. 3, our FreestyleRet framework, adopting 4 prompt tokens, outperforms other number settings including 1, 2, 8 prompt tokens under three evaluation metrics.

Computation Comparison. To validate the lightweight nature of FreestyleRet and its ease of integration into existing retrieval models, we analyze the computational complexity of our framework compared with other baselines. Table. 4 shows the statistical analysis of trainable parameters and inference time per batch for our FreestyleRet framework and other baselines. Compared with the

Table 3: The analysis for the prompt token design. We conduct the ablations for the prompt tokens' insert position, the initialization feature, and the prompt token number in our FreestyleRet framework to demonstrate the performance impact of each component on style-diversified retrieval.

#	Shallow	Bottom	Token-Num	Sketch→Image	Art→Image	Low-Res→Image
1	Random	Random	4	68.1	63.5	78.8
2	Style Space	Random	4	76.7	69.1	82.4
3	Random	Gram	4	76.8	69.2	81.8
4	Gram	Style Space	4	78.1	69.5	84.4
5	Style Space	Gram	4	**81.6**	**72.3**	**86.7**
6	Style Space	Gram	1	68.2	64.7	79.1
7	Style Space	Gram	2	72.3	65.9	82.8
8	Style Space	Gram	8	77.9	67.1	80.7

Table 4: Comparison of the computation complexity. Our framework is computationally efficient from the trainable parameter and inference speed aspects.

Method	Parameters(M)	Speed(ms)
CLIP	427M	68ms
BLIP	891M	62ms
VPT	428M	73ms
FreestyleRet-CLIP	476M$_{(+29)}$	96ms$_{(+28)}$
FreestyleRet-BLIP	940M$_{(+29)}$	101ms$_{(+39)}$

Table 5: Evaluate on FSCOCO dataset for the sketch retrieval task.

Method	R@1	R@10
QST	23.6	52.9
SCM	23.4	52.6
CrossAttn	23.7	53.5
SceneTrilogy	24.1	53.9
FreestyleRet	**29.6**	**56.1**

multi-modality models, our FreestyleRet is lightweight both in the trainable parameter and the inference speed. Compared with the cross-modality models, including CLIP and BLIP, our framework slightly increases the inference time and the trainable parameter while maintaining rapid deployment and application without significant impact.

Performance on Sketch-Retrieval Task. We additionally evaluate our FreestyleRet on the Sketch Retrieval task, which aims to retrieve images using sketch queries. We apply FSCOCO [6] as the evaluation dataset and select recent models as the baselines, including QST [34], SCM [1], CrossAttn, and SceneTrilogy [5]. The baselines and our model utilize CLIP as the visual encoder. As shown in Tab. 5, our FreestyleRet, incorporating style feature extraction and style-init prompt tuning, significantly outperforms other baselines in the sketch retrieval task.

5.4 Qualitative Analysis

In this section, we do the qualitative analysis of our framework's performance by visualizing the high-dimensional feature distribution and the prediction cases from our FreestyleRet-CLIP framework compared with the baseline, the finetuning form of the CLIP model.

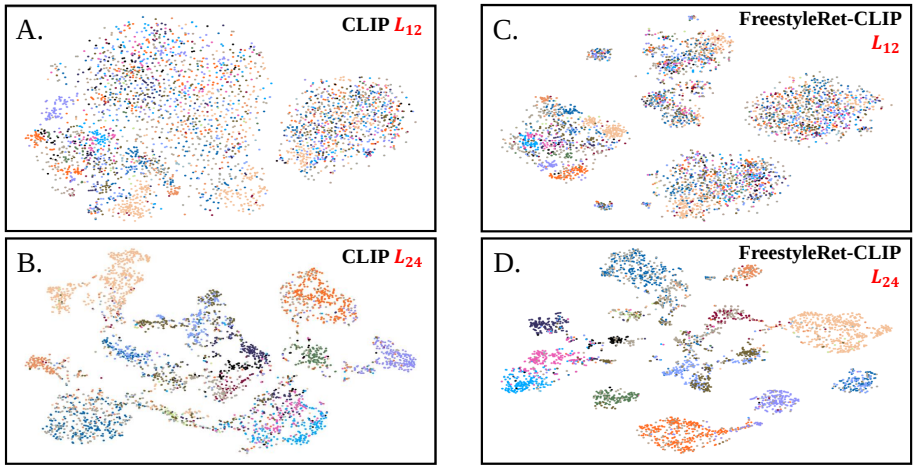

Fig. 5: The Feature Distribution Analysis for our FreestyleRet.

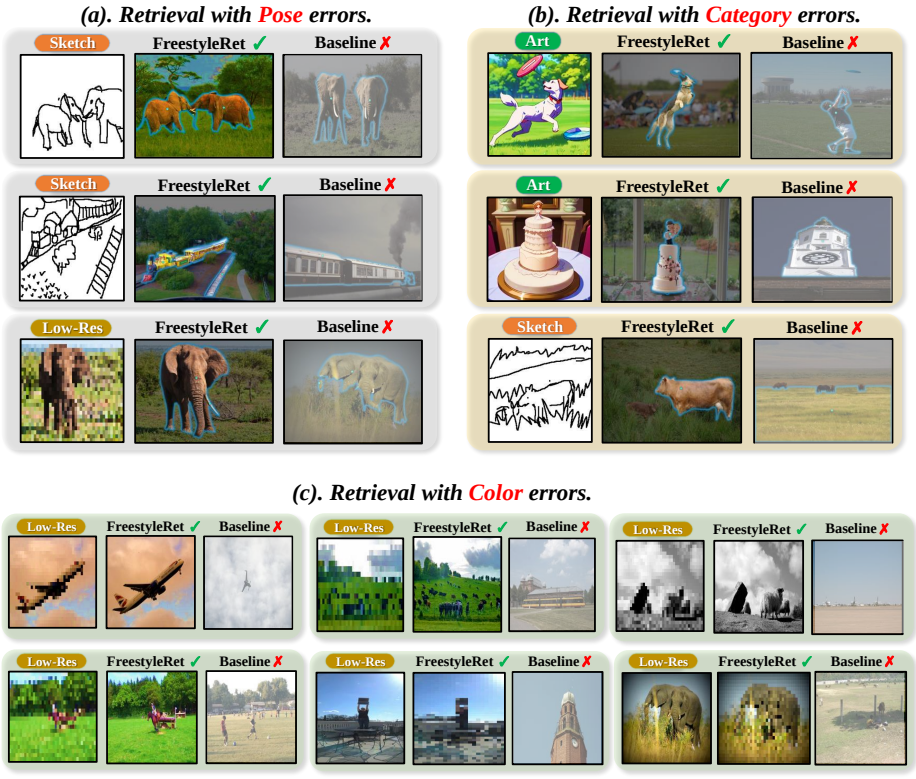

Fig. 6: The Case Study for our FreestyleRet and the CLIP baseline. We visualize style-diversified queries and their corresponding retrieval answers from our FreestyleRet model and the CLIP baseline model.

Feature Distribution Analysis In Fig. 5, we make T-SNE [37] visualization for the middle layer L_{12} and the output layer L_{24} from the FreestyleRet and the CLIP baseline. The colorful dots donate query features, different colors represent different semantics. Compared to the subfigure A-C and B-D, our FreestyleRet achieves better semantic clustering than the CLIP baseline at both middle and deep layers of the transformer.

Case Study and Error Analysis In Fig. 6, we visualize the style-diversified query inputs and their corresponding retrieval answers from our FreestyleRet model and the CLIP baseline model. We summarize three common retrieval errors in the case analysis, where pose errors, category errors, and color errors represent the false retrieval result with false poses, categories, and colors. We propose the pose error cases in Fig. 6(a). The pose information is contained widely in different style queries. Thus, pose error cases occur in sketch, art, and low-resolution queries. The art queries tend to reshape the category into the art form. Thus, in Fig. 6(b), most of the category errors occur in the art-style retrieval task. For the low-resolution query retrieval task, color is vital retrieval information. In Fig. 6(c), we show the color errors from the low-resolution retrieval task. Compared with the CLIP baseline model, our FreestyleRet framework can achieve fine-grained retrieval based on the pose, category, and color information from style-diversified query inputs, demonstrating the superiority of our FreestyleRet framework.

6 Conclusion

In this paper, we are the first to propose the style-diversified query-based image retrieval task to address the issue of limited query style adaptability in current retrieval models. We construct a corresponding dataset, the Diverse-Style Retrieval dataset, for the style-diversified QBIR task. We further propose a lightweight plug-and-play framework, FreestyleRet, to retrieve from style-diversified query inputs. Our FreestyleRet extracts the query's textural and style features from the gram matrix as the style-diversified initialization for the prompt tuning stage. This facilitates the framework in adapting to the style-diversified query-based image retrieval task. Experiment results on the DSR dataset and the ImageNet-X dataset show the effectiveness and computational efficiency of our FreestyleRet framework. In future work, we will incorporate a broader range of query styles into our Diversified-Style Dataset and explore more efficient style-based prompt-tuning strategies for our framework.

Acknowledgements

Hao Li and Yanhao Jia are equal contributions. Li Yuan is the corresponding author. This work was supported by Natural Science Foundation of China (No. 62202014), and Shenzhen Basic Research Program (No.JCYJ20220813151736001)

References

1. Aytar, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., Torralba, A.: Cross-modal scene networks. IEEE transactions on pattern analysis and machine intelligence **40**(10), 2303–2314 (2017)
2. Bossett, D., Heimowitz, D., Jadhav, N., Johnson, L., Singh, A., Zheng, H., Dasgupta, S.: Emotion-based style transfer on visual art using gram matrices. In: 2021 IEEE MIT Undergraduate Research Technology Conference (URTC). pp. 1–5. IEEE (2021)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Cheng, X., Zhang, N., Yu, J., Wang, Y., Li, G., Zhang, J.: Null-space diffusion sampling for zero-shot point cloud completion
5. Chowdhury, P.N., Bhunia, A.K., Sain, A., Koley, S., Xiang, T., Song, Y.Z.: Scenetrilogy: On human scene-sketch and its complementarity with photo and text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10972–10983 (2023)
6. Chowdhury, P.N., Sain, A., Bhunia, A.K., Xiang, T., Gryaditskaya, Y., Song, Y.Z.: Fs-coco: Towards understanding of freehand sketches of common objects in context. In: European Conference on Computer Vision. pp. 253–270. Springer (2022)
7. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (Csur) **40**(2), 1–60 (2008)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R.: A brief review of domain adaptation. Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020 pp. 877–894 (2021)
11. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (June 2016)
13. Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A.: Recommendation systems: Principles, methods and evaluation. Egyptian informatics journal **16**(3), 261–273 (2015)
14. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision (ECCV) (2022)
15. Jin, P., Li, H., Cheng, Z., Huang, J., Wang, Z., Yuan, L., Liu, C., Chen, J.: Text-video retrieval with disentangled conceptualization and set-to-set alignment. arXiv preprint arXiv:2305.12218 (2023)
16. Jin, P., Li, H., Cheng, Z., Li, K., Ji, X., Liu, C., Yuan, L., Chen, J.: Diffusionret: Generative text-video retrieval with diffusion model. arXiv preprint arXiv:2303.09867 (2023)

17. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3668–3678 (2015)
18. Kafai, M., Eshghi, K., Bhanu, B.: Discrete cosine transform locality-sensitive hashes for face retrieval. IEEE Transactions on multimedia **16**(4), 1090–1103 (2014)
19. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)
20. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)
21. Li, H., Huang, J., Jin, P., Song, G., Wu, Q., Chen, J.: Weakly-supervised 3d spatial reasoning for text-based visual question answering. IEEE Transactions on Image Processing (2023)
22. Li, H., Li, X., Karimi, B., Chen, J., Sun, M.: Joint learning of object graph and relation graph for visual question answering. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 01–06. IEEE (2022)
23. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
24. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
25. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021)
26. Li, X., Yang, J., Ma, J.: Recent developments of content-based image retrieval (cbir). Neurocomputing **452**, 675–689 (2021)
27. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. Advances in neural information processing systems **30** (2017)
28. Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., Tang, J.: P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 61–68 (2022)
29. Liu, X., Ji, K., Fu, Y., Tam, W.L., Du, Z., Yang, Z., Tang, J.: P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602 (2021)
30. Liu, Y., Lu, Y., Liu, H., An, Y., Xu, Z., Yao, Z., Zhang, B., Xiong, Z., Gui, C.: Hierarchical prompt learning for multi-task learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10888–10898 (2023)
31. Meng, X., Huang, J., Li, Z., Wang, C., Teng, S., Grau, A.: Dedustgan: Unpaired learning for image dedusting based on retinex with gans. Expert Systems with Applications **243**, 122844 (2024)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
34. Song, J., Song, Y.Z., Xiang, T., Hospedales, T.M.: Fine-grained image retrieval: the text/sketch input dilemma. In: BMVC. vol. 2, p. 7 (2017)

35. Tao, Y.: Image style transfer based on vgg neural network model. In: 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). pp. 1475–1482. IEEE (2022)
36. Thomee, B., Lew, M.S.: Interactive search in image retrieval: a survey. International Journal of Multimedia Information Retrieval **1**, 71–86 (2012)
37. Van Der Maaten, L.: Learning a parametric embedding by preserving local structure. In: Artificial intelligence and statistics. pp. 384–391. PMLR (2009)
38. Wang, P., Li, Y., Vasconcelos, N.: Rethinking and improving the robustness of image style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 124–133 (2021)
39. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. arXiv preprint arXiv:2212.00490 (2022)
40. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
41. Xu, S., Pang, L., Shen, H., Cheng, X.: Match-prompt: Improving multi-task generalization ability for neural text matching via prompt learning. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 2290–2300 (2022)
42. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. arXiv preprint arXiv:2303.09833 (2023)
43. Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering **34**(12), 5586–5609 (2021)
44. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
45. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
46. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)