Training-free Conditional Image Embedding Framework Leveraging Large Vision Language Models

Supplementary Material

A. Datasets

Tab. A-1 shows the datasets used in our experiments with Language-Guided Zero-Shot Deep Metric Learning (**LanZ-DML**) [20], **Style Similarity** [47], and General Conditional Image Similarity (**GeneCIS**) [53].

LanZ-DML The LanZ-DML task includes five datasets: Synthetic Cars, Cars196, CUB200, DeepFashion, and the Movie Poster dataset. The Synthetic Cars dataset consists of artificially generated images of cars on solid-colored backgrounds. The conditions for this dataset include the car model, background color, and car color. Cars196 is also a dataset related to cars, but unlike the Synthetic Cars dataset, it contains real images of cars along with their actual backgrounds. The Cars196 dataset, unlike the Synthetic Cars Datasets, contains images of real cars. In previous studies, experiments were conducted using the Cars196 dataset with three conditions: car model, manufacturer, and car type. However, since the metadata was not available, this experiment was conducted using only the car model as the condition. CUB200 is a dataset that contains 200 different bird species. The only condition for this dataset is the bird species. Following the approach of previous studies, 100 species were selected from the 200 available in the dataset for the experiment. The DeepFashion dataset is a clothing dataset, and it includes four conditions: Clothing Category, Texture, Fabric, and Fit, making it the most diverse in terms of conditions. The Movie Poster dataset consists of movie posters and includes two conditions: movie genre and production country. For all these datasets, including Synthetic Cars, Cars196, CUB200, DeepFashion, and Movie Poster, there is no separate index image, so the query image itself is used as the index image for retrieval.

Style Similarity The Style Similarity task includes two datasets: DomainNet and WikiArt. Wikiart is a dataset focused on image styles and includes a wide variety of artistic styles. The number of different styles in this dataset is 1,119, making it the most diverse among the datasets used. Separate index images are provided, with a total of 16,006 query images and 64,090 index images. DomainNet is a dataset that includes images from various domains and features two conditions: Style and Object. The Style condition includes six types: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. However, since Quickdraw is very similar to Sketch, Quickdraw is excluded from experiments, following the approach of previous studies. The Object condition consists of 345 categories, including airplane, train, and dog. Domain Net also has index images separate from the query images, with the total number of index images being 238,087. While previous studies focused solely on style, we conduct experiments that also consider the object.[1]

GeneCIS GeneCIS benchmark, unlike other datasets, includes two types of tasks: focus and change. The focus task, similar to other datasets, involves obtaining embeddings that highlight features or objects present in the images. In contrast, The change task involves creating embeddings for synthesized images that highlight specific features according to given conditions. Each task—focus and change—includes two conditions: object and attribute. In our experiments, we perform evaluations for the focus task under two conditions: attribute and focus. For each query image in the GeneCIS benchmark, a set of 10 or 15 index images is provided, of which only one is the correct match. The evaluation measures the system's ability to accurately retrieve the correct image. Furthermore, since the GeneCIS dataset includes different conditions for each query image, the number of condition classes corresponds to the number of query images.

B. LanZ-DML Benchmark Experiments

InDiReCT Baseline InDiReCT extract features from CLIP's image embeddings using a PCA-like approach with a predefined list of labels as conditions. However, since obtaining such a label list is not feasible in practice, we generate labels using a LLM. Specifically, we use GPT-4.1 with the following prompt to generate a list of labels: "`Generate {num_generate} labels in English to classify {condition} and output them in comma-separated format:`" Here, `{num_generate}` represents the number of labels to generate. We generate labels in increments of 10, ranging from 10 to 100. Fig. B-1 presents the MAR@R scores for different numbers of labels, and we report the maximum value among them. We denote the case where the correct labels from the dataset are used as InDiReCT (Oracle), while the case where the generated labels are used is reported as InDiReCT (Real).

[1] Since the dataset split from previous studies was not available, a new subset was created for this experiment, consisting of 10,000 images as query images and 100,000 images as index images. This split will be publicly available on GitHub.

Table A-1. Statistics of the datasets used for each task: LanZ-DML and Style Similarity GeneCIS. * indicates the number of index images per query image.

Task	Dataset	#Query Image	#Index Image	Condition	#Condition Class
LanZ-DML	Synthetic Cars	1,000	-	Car Model Car Color Background Color	6 18 18
	Cars196	8,131	-	Car Model	98
	CUB200	4,462	-	Bird Species	100
	DeepFashion	4,000	-	Clothing Category Texture Fabric Fit	50 7 6 3
	Movie Posters	8052	-	Genre Production Country	25 69
Style Similarity	Wikiart	16,006	64,090	Style	1,119
	DomainNet	10,000	100,000	Object Style	345 5
GeneCIS	GeneCIS benchmark	2,000 1,960	10* 15*	Attribute Object	2,000 1,960

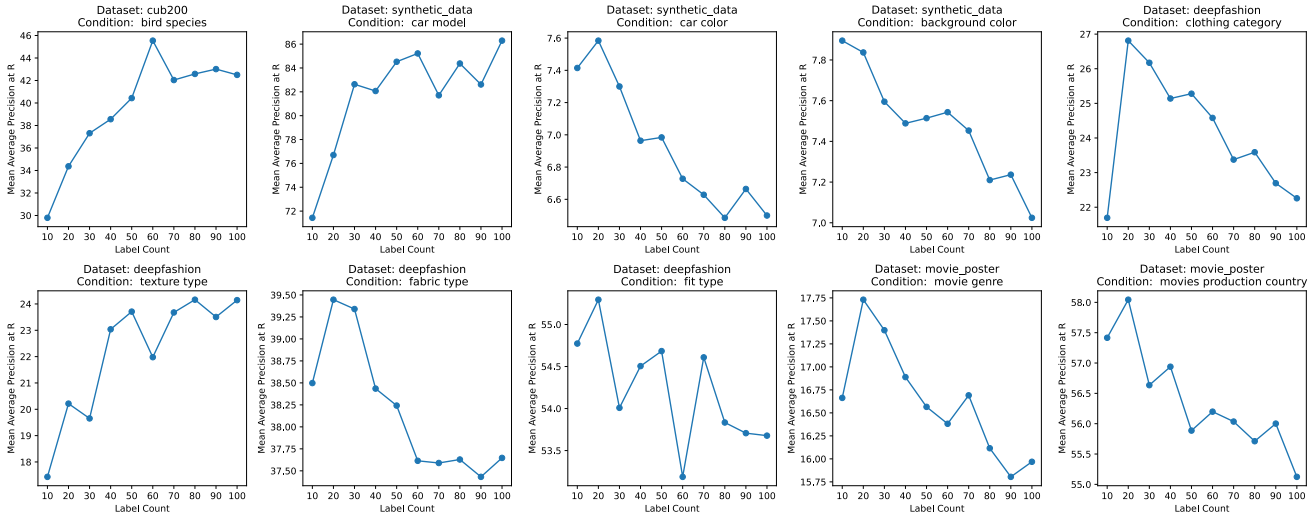

Figure B-1. The MAP@R score for each number of generated labels in InDiReCT.

BLIP/BLIP-2 Baselines To further contextualize the performance of our method, we included two additional vision-language models, BLIP and BLIP-2, as baselines in our LanZ-DML benchmark experiments. For the BLIP baseline, we used the "`blip-itm-large-coco`" model. The image was processed by the vision encoder, while the textual condition was fed into the text encoder. We then extracted the `[CLS]` token from the final multimodal output to serve as the conditional image embedding. For BLIP-2, we employed the "`blip2-flan-t5-xl`" model. In this setup, the image was processed by the vision encoder while the textual condition was passed to the text encoder, and we then used the averaged output from the Q-Former as the final embedding.

As shown in the full results in B-1, both BLIP and BLIP-2 exhibit significantly lower performance on this conditional retrieval task compared to CLIP-based methods and DIOR. This outcome is likely due to fundamental differ-

Table B-1. Evaluation results for the LanZ-DML benchmark. Since we could not obtain the text labels from the metadata of Cars196, they are marked as N/A in InDiReCT. * indicates that the vision encoder is not CLIP-L/14, † denotes methods that require training, and ‡ denotes methods that require prior knowledge of Oracle labels included in each condition. Avg represents the average over the respective conditions for the Synthetic Cars, Cars196 (Cars), CUB200 (CUB), DeepFashion, and Movie Posters datasets.

		Synthetic Cars			Cars	CUB	DeepFashion				Movie Posters		
	Params	Mod.	Col.	BG	Mod.	Spe.	Clo.	Tex.	Fab.	Fit	Gen.	Cou.	Avg
CLIP	428M	72.3	5.8	6.8	43.0	33.3	12.1	18.4	33.9	53.0	12.2	50.7	31.0
EVA-CLIP*	18B	91.9	5.9	6.8	67.7	60.2	15.3	18.8	35.9	53.5	10.1	49.7	37.8
BLIP	450M	22.5	10.2	7.6	9.9	15.5	11.1	14.5	33.3	54.3	7.3	48.3	19.5
BLIP2	4B	42.2	5.2	7.7	27.7	6.1	11.7	16.9	35.3	54.3	8.0	51.3	24.2
InDiReCT (CLIP)†‡	428M	94.6	7.3	7.8	N/A	44.0	24.9	24.9	40.1	54.4	17.9	55.6	37.1
InDiReCT (CLIP)†	428M	86.3	7.6	7.9	N/A	45.5	26.8	24.2	39.4	55.3	17.7	58.0	36.9
InDiReCT (EVA-CLIP)*†‡	18B	96.7	7.7	8.0	N/A	67.1	28.9	24.9	41.5	55.0	15.2	55.2	40.0
InDiReCT (EVA-CLIP)*†	18B	**94.0**	7.5	7.8	N/A	**66.9**	32.1	24.0	40.6	54.9	14.6	52.3	39.5
CapEmb (SimCSE)	11B	51.2	11.5	11.5	28.6	24.4	23.0	13.9	39.4	52.4	18.3	53.6	29.8
CapEmb (SentenceT5)	11B	49.6	14.8	12.7	24.8	23.1	25.7	14.6	39.2	52.4	17.9	49.9	29.5
DIOR (LLaVA-1.6)	7B	73.7	13.9	15.6	40.6	21.7	26.7	26.6	42.8	54.5	20.3	58.1	35.9
DIOR (LLaVA-1.6)	13B	78.9	14.9	15.4	45.8	23.1	30.8	**34.8**	43.4	57.4	21.9	60.0	38.8
DIOR (Llama-3.2)	11B	88.0	14.7	**17.8**	67.3	51.9	33.7	26.0	46.4	57.8	23.4	62.2	44.5
DIOR (Qwen-2.5-VL)*	3B	71.0	14.4	15.3	65.5	57.7	42.7	31.2	43.6	58.4	21.5	61.1	43.9
DIOR (Qwen-2.5-VL)*	7B	73.2	**15.0**	16.8	**68.0**	58.1	**45.3**	33.9	**47.6**	**62.1**	**28.1**	**66.0**	**46.7**

ences in their training objectives and architecture. BLIP and BLIP-2 are primarily optimized for image-text matching (ITM), a task that predicts whether an image-text pair is correctly matched. This objective does not inherently cultivate a well-structured cosine similarity space for nuanced image-to-image comparisons, placing them at a disadvantage against models like CLIP. Furthermore, the Q-Former in BLIP-2 is designed to produce a compact representation to interface with a Large Language Model (LLM), which may result in a lower-resolution embedding that is less effective for standalone similarity tasks.

C. GeneCIS Experiments

Settings GeneCIS [53] is a benchmark that retrieves images that match the query image and condition, with objects and attribute types as important dimensions for image similarity. Although this includes two types of tasks, including "focus" and "change," we work on the "focus" task, which is consistent with our goals. The benchmark is built on existing public datasets, such as COCO [18, 26] and VAW [40] based on Visual Genome [23]. For each query, 10 to 15 images are included, one of which is the correct answer, and the evaluation metric is Recall at k (Recall@k) for k=1, 2, and 3. For baselines, we compare against a Combiner model [3], which is trained with contrastive learning to combine image and text information. We report results from two versions of the Combiner trained on different datasets: one on CIRR [30] and another on CC3M [46]. Furthermore, following the prior study, we include a CLIP baseline with three variants: (1) using the embedding of the text condition from CLIP text encoder (Text-only), (2) using the embedding of the query image from CLIP image encoder (Image-only), and (3) using an averaged embedding of them (Text+Image).

Table C-1. Evaluation Results on the GeneCIS benchmark.

	Focus Attribute (Recall@k)			Focus Object (Recall@k)		
Method	1	2	3	1	2	3
CLIP (Text only)	10.2	20.5	29.6	6.5	16.8	22.4
CLIP (Image only)	17.7	30.9	41.9	9.3	18.2	26.2
CLIP (Text + Image)	15.6	26.3	37.1	10.8	21.0	31.2
Combiner (CIRR)	15.1	27.7	39.8	13.5	25.4	36.7
Combiner (CC3M)	19.0	31.0	41.5	14.7	25.9	36.1
DIOR	**24.0**	**36.8**	**47.0**	**21.1**	**34.3**	**42.9**

Results We compare DIOR against methods that require task-specific training on the GeneCIS benchmark. As shown in Tab. C-1, DIOR achieves strong performance despite not being trained specifically for this task. This suc-

Table D-1. Results of text generation using DIOR prompts on DomainNet and WikiArt. GT stands for Ground Truth.

	Image	Model Output	GT
DomainNet		`Describe the image in one word regarding the object:` **`dog`**	dog
Wikiart		`Describe the image in one word regarding the name of the artist:` **`Childe Hassam`**	George Luks

cess is likely due to the effective use of the LVLM's inherent knowledge to correctly reference attributes and objects within the images.

D. Analysis

Performance Degradation in Highly Specific Domains To investigate the performance degradation observed specifically in the Wikiart dataset, as shown in Tab. D-1. For the DomainNet dataset, the output was "dog," correctly matching the ground truth. However, for the Wikiart dataset, "*Childe Hassam*" is simply the wrong output. We can at least confirm that the decoded tokens are not corrupted, but the model confidently provides the wrong answer. We attribute this failure to the LVLM's lack of profound knowledge in a specific domain. Since LVLMs are applied across various tasks, visual instruction tuning involves training with images and responses from various domains. DomainNet, with its diverse image types, closely aligns with the photos typically used in visual instruction tuning. In contrast, Wikiart represents a specialized art domain that demands deep, nuanced knowledge for accurate image interpretation. We conjecture that our base LVLM may lose such knowledge during construction. Additionally, fine-tuning LVLMs for particular domains is a promising approach, which we will leave for future research.

Evaluation on Image Classification Tasks To further assess the quality of the generated, we conducted additional experiments on standard image classification tasks. We evaluated the performance of our embeddings on linear probing and few-shot classification using the Cars196 and CUB200 datasets, featuring a single condition ("Car Model" and "Bird Species," respectively). For this evaluation, we compare three methods: (1) the CLIP ViT-L/14 baseline, (2) our proposed DIOR framework with the conditional prompt using Llama-3.2 (referred to as DIOR (Ours)), and (3) an ablation variant of DIOR that uses a non-conditional prompt ("Describe the image in one word:") to investigate the impact of conditioning, referred to as DIOR (w/o condition).

Table D-2. Linear probe and few-shot classification accuracy (%) on the Cars196 and CUB200 datasets. Our proposed method, DIOR (Ours), significantly outperforms both the CLIP baseline and its non-conditional variant, highlighting the effectiveness of our conditional prompting strategy. Best results are in **bold**.

Cars196	**Linear Probe**			**Few-shot**		
Method	k=1	k=5	k=10	k=1	k=5	k=10
CLIP ViT-L/14	59.82	79.41	85.38	55.41	81.18	86.04
DIOR (w/o condition)	40.06	58.83	69.53	39.44	64.14	72.43
DIOR (Ours)	**80.02**	**89.67**	**93.15**	**77.47**	**91.33**	**93.38**

CUB200	**Linear Probe**			**Few-shot**		
Method	k=1	k=5	k=10	k=1	k=5	k=10
CLIP ViT-L/14	49.11	68.80	75.59	46.15	72.00	78.42
DIOR (w/o condition)	31.28	46.41	52.78	33.22	53.77	60.66
DIOR (Ours)	**62.21**	**75.57**	**78.94**	**60.89**	**79.54**	**81.81**

The results are presented in Tab. D-2. Our proposed DIOR method consistently and significantly outperforms the strong CLIP baseline across all metrics in both linear probe and few-shot settings. For instance, on Cars196, DIOR achieves a linear probe accuracy of 80.02% (k=1), a substantial improvement over CLIP's 59.82%. This demonstrates that the conditional embeddings generated by DIOR are not only effective for retrieval but also produce a more linearly separable feature space, indicating higher-quality representations.

Crucially, the performance of DIOR without the conditional prompt (DIOR w/o condition) drops significantly, falling even below the CLIP baseline. This result strongly validates our core hypothesis: the performance gain is not merely due to using a larger LVLM, but is fundamentally driven by the *conditional* nature of our prompting strategy. This highlights the effectiveness of the DIOR framework in steering the LVLM to focus on specific, relevant attributes and generate highly discriminative embeddings.

E. Case Study

Retrieval results on DeepFashion obtained through CLIP and DIOR are shown in Fig. E-1. CLIP can find the correct image similar to the query image for the clothing category, but it fails for the texture type. This suggests that image embeddings of CLIP are primarily focused on the clothing category features. In contrast, DIOR demonstrates appropriate retrieval in both condition. This highlights potential advantages of DIOR in tasks that require more detailed feature distinction.

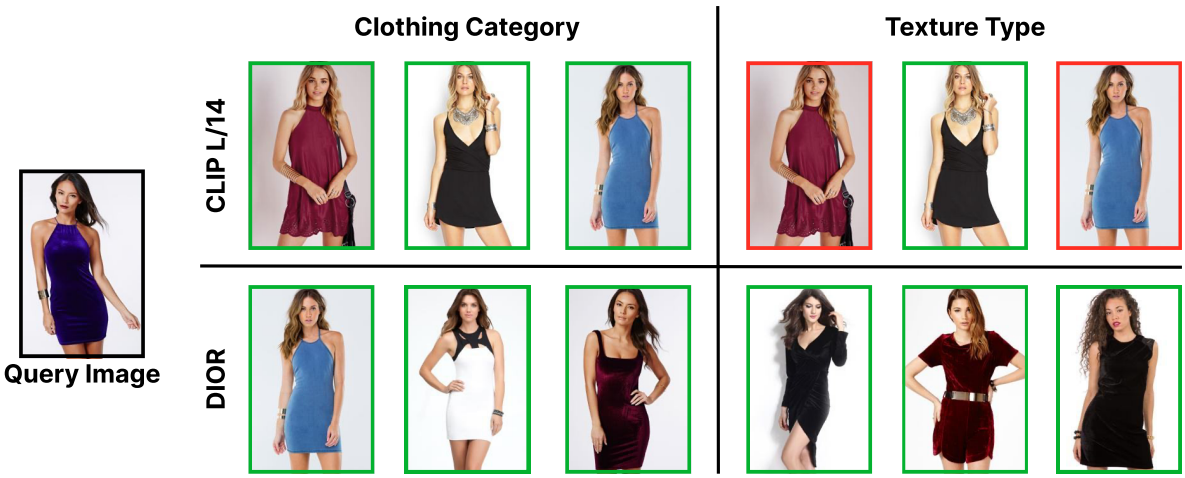

Figure E-1. Examples of images retrieved by CLIP L/14 and DIOR in terms of 'Clothing Category' and 'Texture Type' for the query image in DeeepFashion dataset. Among the retrieved images, the ones with green boxes indicate correct cases and the ones with red boxes indicate incorrect cases.