CLAY: Conditional Visual Similarity Modulation in Vision-Language Embedding Space

Supplementary Material

Contents

In this supplementary material, we provide more details and results which are not included in the main paper due to space constraints. Sec. A details of our method with pseudo code. In Sec. B, we provide the construction process of our synthetic dataset CLAY-EVAL. Sec. C offers experimental details including setup and implementations. We provide additional results of ablation, qualitative, t-SNE in Sec. D. Finally, we discuss limitations of our method in Sec. E. We further kindly encourage readers to see the supplementary HTML file containing additional qualitative results.

A. Method Details

In this section, we provide a detailed explanation of our method. Our method consists of two main processes: (1) construction of projection matrix and (2) computation of visual similarity between query and database images.

Construction of projection matrix. As presented in the main paper, we first generate condition-related text prompts with Large Language Model (LLM). We automatically construct a system prompt template with LLM in advance, to effectively filter non-related generations and ensure the response template, as shown in Tab. S1. Then, we input the LLM to generate texts with given condition c, followed by forwarding step to the text encoder of VLM:

$$c_i \in \mathcal{C}, \quad i = 1, \dots, n, \tag{1}$$

$$\mathbf{t}_i^c = f_T(c_i), \tag{2}$$

$$\mathbf{T}_c = [\mathbf{t}_1^c, \mathbf{t}_2^c, \dots, \mathbf{t}_n^c]^\top, \tag{3}$$

where c_i is the i-th LLM generated text prompt. As in Sec. 4, we map the features onto the tangent space before applying Singular Value Decomposition (SVD). We choose the normalized mean of text features as a reference point, and define *logarithm map* to map the points lying on the hyperspherical manifold onto the tangent space of $\boldsymbol{\mu}_c$ following [3, 8, 10]:

$$\boldsymbol{\mu}_c = \frac{\frac{1}{n}\sum_{i=1}^n \mathbf{t}_i}{\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{t}_i\right\|}, \tag{4}$$

$$\log_{\boldsymbol{\mu}_c}(\mathbf{x}) := (\mathbf{x} - \boldsymbol{\mu}_c(\mathbf{x}^\top \boldsymbol{\mu}_c))\frac{\theta}{\sin(\theta)}, \quad \theta = \arccos(\mathbf{x}^\top \boldsymbol{\mu}_c). \tag{5}$$

After we map the features onto the tangent space with Eq. 5, we then generate projection matrix $\mathbf{P}_c$ by applying SVD:

$$\log_{\boldsymbol{\mu}_c}(\mathbf{T}_c) = \left[\log_{\boldsymbol{\mu}_c}(\mathbf{t}_1^c), \cdots, \log_{\boldsymbol{\mu}_c}(\mathbf{t}_n^c)\right]^\top, \quad \log_{\boldsymbol{\mu}_c}(\mathbf{T}_c) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top, \tag{6}$$

$$\mathbf{P}_c = \mathbf{V}_k\mathbf{V}_k^\top. \tag{7}$$

Computation of visual similarity. To alleviate the approximation error using the *logarithm map* resulting from the gap between text-visual features, also known as *conic effect* [17], we first mitigate the gap by aligning the mean of visual features with $\boldsymbol{\mu}_c$. To achieve this, we apply two householder transformations so that the distances and angles between two visual features remain unchanged, while the visual mean becomes aligned with the textual mean:

$$\tilde{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}_v + \boldsymbol{\mu}_c}{\|\boldsymbol{\mu}_v + \boldsymbol{\mu}_c\|}, \tag{8}$$

$$\mathbf{H}_1 = \mathbf{I} - 2\frac{(\boldsymbol{\mu}_v - \tilde{\boldsymbol{\mu}})(\boldsymbol{\mu}_v - \tilde{\boldsymbol{\mu}})^\top}{\|\boldsymbol{\mu}_v - \tilde{\boldsymbol{\mu}}\|^2}, \quad \mathbf{H}_2 = \mathbf{I} - 2\frac{(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_c)(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_c)^\top}{\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_c\|^2}, \quad H(\mathbf{v}) := \mathbf{H}_2\mathbf{H}_1\mathbf{v}. \tag{9}$$

Therefore, we finally compute the conditional similarity $\texttt{csim}_{\text{CLAY}}$ between *query* image feature $\mathbf{v}_q = f_I(I_q)$ and *database* image feature $\mathbf{v}_d = f_I(I_d)$ by applying $H(\cdot)$, logarithmic map $\log_{\boldsymbol{\mu}_c}(\mathbf{T}_c)$, and projection matrix $\mathbf{P}_c$ as shown in Alg. 1. Then, the final retrieved images are sorted by these values of conditional similarity.

Table S1. **LLM Prompts for generating condition-related text descriptions.**

LLM system prompts.
You will be given a user prompt. Your job is to determine the user intent (where to focus on in the image) and generate 100 prompts related to the intent, following the format provided. For example, given the following user intent: "I want to focus on the human action in this image." An exemplar answer is: Intent: action Prompts: a photo of running, a photo of jumping, a photo of swimming, ... Output must always be in the exact format: Intent: <intent-word> Prompts: <string>, <string>, ..., <string> - The Prompts list must contain exactly 100 unique entries following the format "a photo of {intent}". - Each string should directly replace the placeholder (e.g., human action) with a real example without inserting any other phrases. - Do not include numbering or bullet points or extra quotes.
Generation example.
LLM input : I want to focus on the human action in this image LLM output: Intent: action Prompts: a photo of running, a photo of jumping, a photo of swimming, a photo of dancing, ...

Algorithm 1 Conditional similarity computation

Require: query feature $\mathbf{v}_q$, database image feature $\mathbf{v}_d$
Require: $\log_{\boldsymbol{\mu}_c}(\cdot)$, $\mathbf{T}_c$, $H(\cdot)$
$\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^\top \leftarrow \text{SVD}(\log_{\boldsymbol{\mu}_c}(\mathbf{T}_c))$
$\mathbf{P}_c \leftarrow \mathbf{V}_k \mathbf{V}_k^\top$
$\mathbf{v}_{q'} \leftarrow \mathbf{P}_c \log_{\boldsymbol{\mu}_c}(H(\mathbf{v}_q))$
$\mathbf{v}_{d'} \leftarrow \mathbf{P}_c \log_{\boldsymbol{\mu}_c}(H(\mathbf{v}_d))$
$\texttt{csim}_{\text{CLAY}} \leftarrow \dfrac{\mathbf{v}_{q'} \cdot \mathbf{v}_{d'}}{\|\mathbf{v}_{q'}\| \|\mathbf{v}_{d'}\|}$
return $\texttt{csim}_{\text{CLAY}}$

B. CLAY-EVAL Construction Pipeline

In this section, we provide detailed construction pipeline of our synthetic dataset, CLAY-EVAL. We first present overview, detail the schema of each CLAY-Object and CLAY-Human, and then present the prompts utilized for generation.

B.1. Overview

Our dataset construction follows a systematic four-stage process designed to ensure high-quality, diverse, and well-controlled naturalistic images.

Data Schema Design. For each entity, we define core attributes (primary query conditions) and diversity attributes (used to enhance sample diversity). This stage also includes applying schema-level filtering to automatically exclude logically contradictory combinations. For example, in the CLAY-Human dataset, this step removes 288 samples for impossible scenarios (*e.g.*, "driving" in "indoor"). Regarding the core and diversity attributes, please refer to Sec. B.2 and B.3.

Prompt Generation. All core and diversity attributes are programmatically combined according to a structured prompt template (detailed in Sec. B.4) to create the final, comprehensive set of text prompts.

Image Generation. All samples are generated using the FLUX.1-dev model [4] with parameters specified in Sec. B.4.

Curation Process. We conduct a rigorous manual review to ensure high data quality. To facilitate this large-scale review efficiently, we utilize a custom HTML-based viewer as illustrated in Fig. S1. This tool is designed to allow for selective sorting and filtering by attributes, enabling us to verify that intended compositions are visually generated correctly. It also provides functionality to select problematic samples and annotate the reasons for exclusion (managed via JSON). This step is crucial for filtering out samples that show severe text-visual misalignment, visual anomalies, or harmful content.

B.2. Object Entity Dataset

Data Schema and Construction. For the *object* entity, we select three core attributes based on their clear vi-

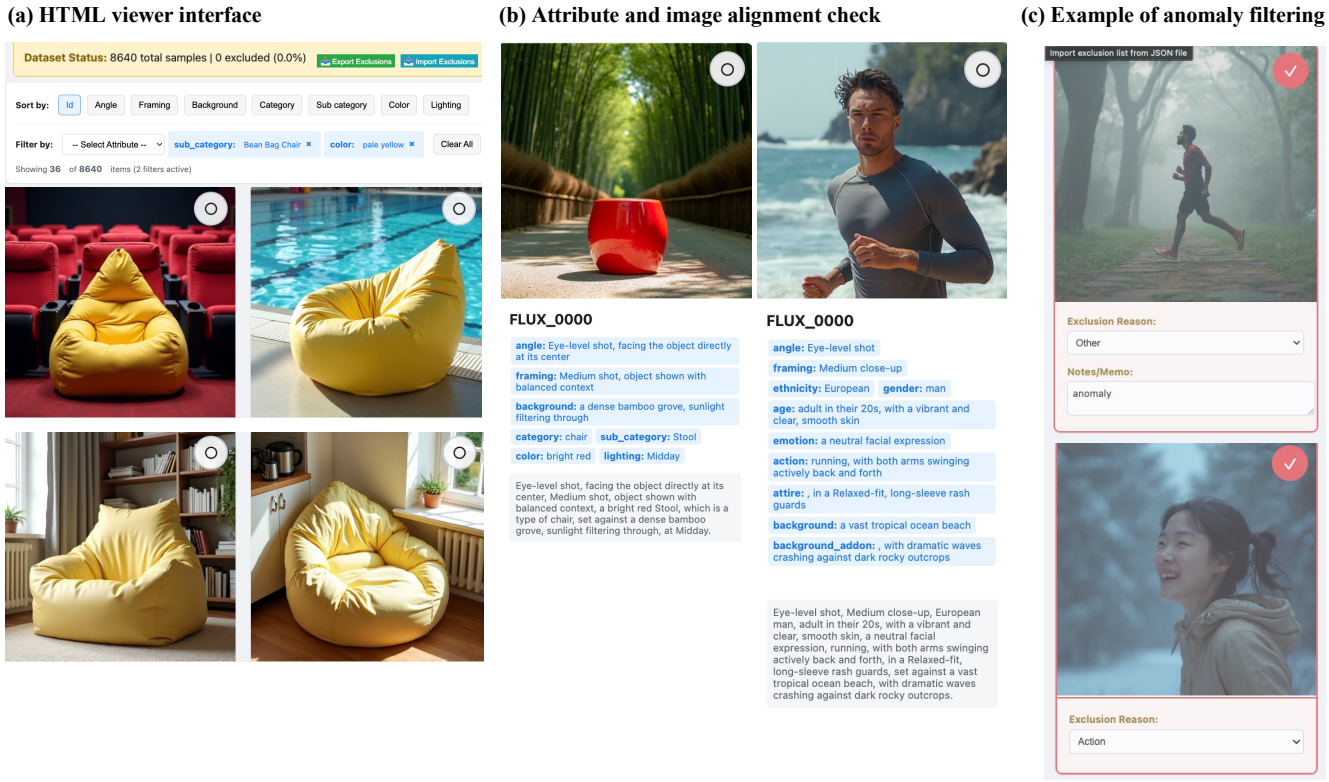


Figure S1. **HTML viewer and sample curation cases.** Illustration of the use of an HTML viewer for quality assurance and data filtering. (a) The overall design and components of the interface used for data inspection. The viewer example shown has been edited for visual clarity. Examples filtered by `sub-category`: Bean Bag Chair and `color`: pale yellow are shown. (b) Verification that the generated image results are correctly aligned with their corresponding attributes (for both *object* and *human* samples) and the full text prompt used for generation. (c) An illustration of the process for filtering out samples exhibiting anomaly or text-visual misalignment. Specific filtering cases are shown: The upper sample was excluded due to a major limb anomaly (artifact). The lower sample was excluded due to textual-visual misalignment, as the specified `action` ("running") was not discernible in the generated image.

sual distinctiveness and hierarchical structure: `category`, `sub-category`, and `color`. To enhance visual diversity, we incorporate three additional diversity attributes: lighting, framing, and angle.

The construction pipeline begins with base combinations of core attributes. As `sub-category` (24 instances) is subordinate to `category` (4 instances), our base combinations are calculated from 10 `colors` and 24 `sub-categories`, resulting in 240 unique combinations (See Tab. S2) . To this set, we add 18 combinations of diversity attributes (3 *lighting* $\times$ 3 *framing* $\times$ 2 *angle*; (See Tab. S3)), yielding 4,320 base combinations (240 $\times$ 18). Finally, we sample uniformly from a pool of 71 distinct *backgrounds* (detailed in Sec. B.5) and generate two samples for each base combination to further increase diversity, resulting in a total of 8,640 images.

Filtering and Final Dataset. During our manual curation process, 1,315 samples with severe anomalies or misalignments are removed. This results in a final dataset of 7,325 samples. The final distributions for core attributes are detailed in Tab. S2 (left column), while qualitative examples per attribute combination are presented in Fig. S2 (a).

B.3. Human Entity Dataset

Data Schema and Construction. For the *human* entity, we define `age`, `action`, and `background` as core attributes, chosen for visual identification. We also introduce five diversity attributes to ensure a balanced and representative dataset: race, gender, emotion, framing, and angle.

The base combinations consist of 3 `age` $\times$ 5 `action` $\times$ 5 `background`, total 75 combinations (See Tab. S2). We then introduce 96 diversity combinations (6 race $\times$ 2 gender $\times$ 2 emotion $\times$ 2 framing $\times$ 2 angle; (See Tab. S3)). This results in 7,200 total combinations (75 $\times$ 96). To increase background detail without changing the total sample count, we prepare two distinct detailed descriptions ("add-ons") for each of the 5 core `backgrounds` (detailed in Sec. B.5) and sample uniformly between them during generation.

Filtering and Final Dataset. Following the schema-level filtering (which removed 288 logical exclusions), 6,912 com-

Table S2. **Prompt Templates and Core Attributes**

	Object Entity	Human Entity
Template	... {color} {sub-category}, which is a type of {category}, ...	... {age}, {action}, ... set against {background} ...
Example	... bright red Stool, which is a type of chair, ...	... adult in their 20s, ... running, with both arms swinging actively back and forth, ... set against a vast tropical ocean beach ...

Core Attributes

Category (4 types)	Vehicle (29.19%), Shoes (28.68%), Bag (21.49%), Chair (20.64%)	**Age** (3 types):	Adult (33.08%), Middle-aged (33.57%), Elderly (33.36%)
Sub-category (24 types)	Backpack (4.83%), Bean Bag Chair (4.82%), Duffel Bag (4.79%), Sneakers (4.76%), Armchair (4.64%), High Heels (4.57%), Tote Bag (4.57%), Sedan (4.57%), Loafers (4.53%), Boots (4.52%), Rocking Chair (4.51%), Van (4.51%), Motorcycle (4.45%), Stool (4.31%), Truck (4.20%), Dress Shoes (4.18%), Suitcase (4.10%), Sandals (3.99%), Bus (3.97%), Bicycle (3.77%), SUV (3.71%), Shoulder Bag (3.19%), Recliner (2.36%), Slippers (2.13%)	**Action** (5 types):	Texting (21.25%), Reading (21.20%), Clapping (21.08%), Running (20.53%), Driving (15.94%)
Color (10 types) :	bright red (10.38%), vivid purple (10.35%), jet black (10.29%), vibrant orange (10.29%), pure white (10.28%), pale yellow (10.21%), deep blue (10.16%), light pink (10.12%), muted gray (9.30%), dark brown (8.63%)	**Background** (5 types):	Beach (21.25%), Snowy (20.80%), Park (20.76%), City (20.58%), Indoor (16.62%)

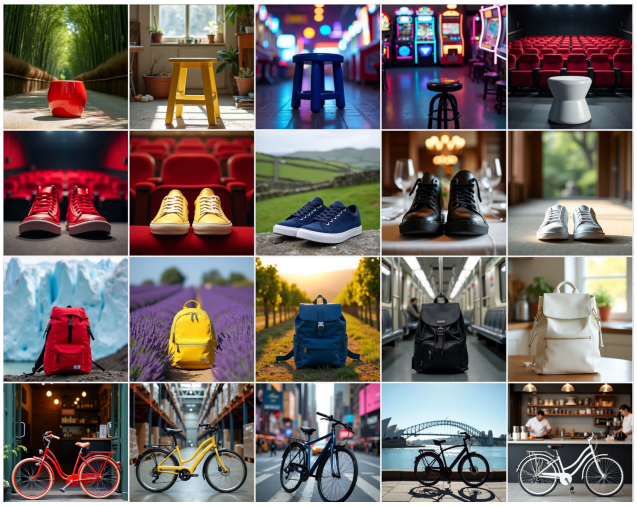
Table S3. **Full Prompt Templates with Diversity Attributes**

	Object Entity	Human Entity
Template	"{angle}, {framing}, a {color} {sub-category}, which is a type of {category}, set against {background}, at {lighting}"	"{angle}, {framing}, {race} {gender}, {age}, {emotion}, {action}, {attire}, set against {background}{background-add-on}"
Example	"Eye-level shot, facing the object directly at its center, Medium shot, object shown with balanced context, a bright red Stool, which is a type of chair, set against a dense bamboo grove, sunlight filtering through, at Midday." (See Fig. S1 (b), left column.)	"Eye-level shot, Medium close-up, European man, adult in their 20s, with a vibrant and clear, smooth skin, a neutral facial expression, running, with both arms swinging actively back and forth, in a Relaxed-fit, long-sleeve rash guards, set against a vast tropical ocean beach, with dramatic waves crashing against dark rocky outcrops." (See Fig. S1 (b), right column.)

Diversity Attributes

Angle (2 types)	Eye-level, High-angle	**Angle** (2 types)	Eye-level, Side profile
Framing (3 types)	Close-up, Medium, Wide	**Framing** (2 types):	Medium close-up, Full shot
Background (71 types)	Urban – Global& Iconic,	**Race** (6 types)	European, Asian, Indian, Middle Eastern, African, Latino
(See Sec. B.5)	Indoor - Residential & Office Indoor - Public & Commercial Indoor - Cultural & Specialized Indoor - Transportation Nature - Specific Flora & Weather Nature - Global Biomes & Landscapes	**Gender** (2 types) **Emotion** (2 types)	Man, Woman Neutral, Joy
Lighting (3 types)	Midday, Morning, Evening		

(a) CLAY-Object



(b) CLAY-Human

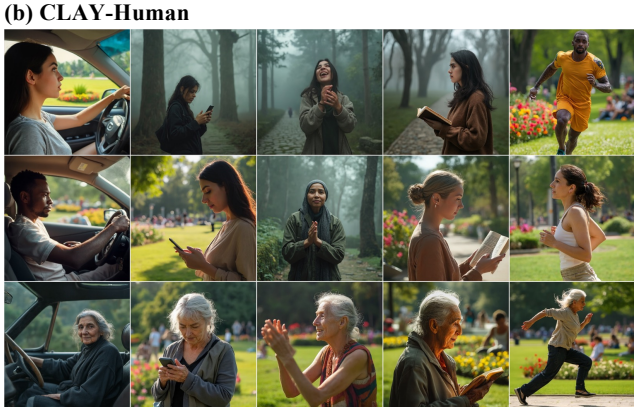


Figure S2. **Diverse naturalistic image samples generated according to intended attribute conditions.** We visualize image samples based on specific *object* entity and *human* entity attributes. (a) The grid is structured by the `Color` attribute (horizontal axis) and the `Category` attribute (vertical axis). Horizontal Axis (`Color`, 5 Columns, Left to Right): Red, Yellow, Blue, Black, White. Vertical Axis (`Category`, 4 Rows, Top to Bottom): Chair (`Sub-category`: Stool), Shoes (`Sub-category`: Sneakers), Bag (`Sub-category`: Backpack), Vehicle (`Sub-category`: Bicycle). (b) The grid is structured by the `Action` attribute (horizontal axis) and the `Age` attribute (vertical axis). Horizontal Axis (`Action`, 5 Columns, Left to Right): Driving, Texting, Clapping, Reading, Running. Vertical Axis (`Age`, 3 Rows, Top to Bottom): Adult, Middle-aged, Elderly.

binations remain. From this set, we manually curate and remove additional 165 samples. This results in a final dataset of 6,745 samples. The final distributions for core attributes are detailed in Tab. S2 (right column), and qualitative examples with each attribute combination are in Fig. S2 (b).

B.4. Generation Details

Prompt Templates. Based on the designed data schema, each text prompt is generated with diverse combinations of attribute instances.

Model and Parameters. We summarize the implementation details of FLUX below.

- **Model ID:** "black-forest-labs/FLUX.1-dev"
- **seed:** 42
- **height:** 512
- **width:** 512
- **guidance_scale:** 0.7
- **num_inference_steps:** 50

Negative Prompt. A negative prompt including the following keywords is used to suppress the generation of unintended or inappropriate images. "nude, naked, underwear, lingerie, topless, ..."

B.5. Detailed Attribute Instances

For visibility, several attribute instances mentioned in the paper are denoted using abbreviations (*e.g.*, in `sub-category`: Rocking Chair $\rightarrow$ Rocking).

Backgrounds for Object Entity. A pool of 71 prompts under 7 major categories is used. The type of each category and the number of contained instances are shown in Tab. S3. Example from Nature - Global Biomes & Landscapes: "rolling green hills and stone walls of the Irish countryside".

Background Add-ons for Human Entity. Two detailed prompts are uniformly sampled for each of the 5 base `background` types. Example from "a vast tropical ocean beach"; – Add-on 1: "with palm trees swaying gently in the tropical breeze" – Add-on 2: "with dramatic waves crashing against dark rocky outcrops".

C. Experimental Details

C.1. Experimental Setup

For evaluation, we consider the real-world fine-grained image classification datasets for single conditional retrieval setting, including Stanford40 [27], OxfordPets [22], PPMI [25, 26], Flowers102 [21], StanfordCars [15], FGVC-Aircraft [18], and Food-101 [5]. We randomly split each dataset into a 1:9 ratio for the construction of *query* and *database*, the final number of images used to evaluation are presented in Tab. S4.

C.2. Implementation Details

We request Large Language Model to generate $n = 100$ prompts, and encode them into the text encoder of Vision Language Model. However, we find that the outputs do not strictly follow the specified total counts. Nevertheless, we utilize the entire set of generated prompts in our experiments. In the process of Singular Value Decomposition, we truncate the top-k singular vectors to obtain projection matrix following previous works [7, 20]. In our experiments, we

Table S4. **Total number of query and database images in evaluation datasets.**

	Action	Mood	Location	Cat	Dog	Instrument	Flower	Car	Aircraft	Food	Clevr4	CLAY-Object	CLAY-Human
Query	937	100	97	235	493	480	775	804	333	2525	1053	732	674
Database	8595	900	903	2136	4485	4320	7414	7237	3000	22725	9478	6593	6071

Table S5. **Text instructions for comparisons across different baseline models.**

Condition	MagicLens	SEARLE	VLM2Vec / QwenVLEmbed / InstructBLIP
Single condition			
Cat, Dog	A {condition} of the same breed	a photo of the same {condition} species of $	What breed of {condition} is in the image?
Car, Aircraft	The same {condition} model	a photo of the same {condition} model of $	What model of {condition} is in the image?
Count	The same object {condition}	a photo of the same object {condition} of $	How many objects are in the image?
Age	The same human {condition}	a photo of the same human {condition} of $	What age of human in the image?
Color, Category (Object)	The same object {condition}	a photo of the same object {condition} of $	What type of object {condition} is in the image?
Etc.	The same (type of) {condition}	a photo of the same (type of) {condition} of $	What type of {condition} is in the image?
Multi condition			
Age, Action	-	-	What age of human and what type of action is in the image?
Age, Background	-	-	What age of human and what type of background is in the image?
Age, Action, Background	-	-	What age of human and what type of action, background is in the image?
Etc.	-	-	What type of {condition1} and {condition2} is in the image?

set $k = 50$. For our comparison methods [2, 6, 16, 19, 28] we set the instructions for describing the text condition as shown in Tab. S5. A few instruction examples in single condition setup are selected by following FocalLens [11], and the remaining samples are chosen manually. In addition, for GeneCIS [24], we set the condition texts as those in the *condition* column of Tab. S5.

D. Additional Results

Backbone variation in VLMs. To demonstrate the effectiveness of our method under large models with ViT-L, we evaluate retrieval performance varying the backbone model on single conditional datasets. Table S6 demonstrates that our method shows substantial gains in mAP compared to the baseline. This suggests that our method has the potential to provide greater improvements when applied to large models.

Comparison on GeneCIS benchmark. We also compare our method with Composed Image Retrieval methods [2, 14, 28] on the GeneCIS benchmark "Focus Attribute" subset. We report Recall at k following previous work [24]. Table S7 shows CLAY still achieves the best performance when utilizing the same or even smaller backbone (*i.e.*, ViT-B).

Multi-conditional retrieval on real dataset. We report multi-conditioned retrieval results of GeneCIS and our method on Stanford40 dataset, using samples labeled with multiple attribute pairs. While GeneCIS cannot input multi-conditioned inputs due to its training setup, we nevertheless applied such inputs for experimental evaluation. In Tab. S8, we find that CLAY achieves higher retrieval performance than GeneCIS, highlighting the effectiveness of our conditioning process on multiple inputs.

Scree plot of singular values. We provide scree plot of singular values in Fig. S3. The top-50 singular values explain over 80% of the variance, implying that the condition-related features can be approximated by our design.

Additional t-SNE plots. We visualize additional t-SNE results of base model (CLIP-B) and ours (CLIP-B) with *database* features in evaluation datasets. Figure S4 shows the representations under varying conditions in CLAY-EVAL-Object dataset, highlighting the effectiveness of our method for separating the conditional visual features. Furthermore, in Fig. S5 and S6, our method can adaptively generate well-separated and structured clusters compared to the base model. Meanwhile, Fig S6 (b) shows the representations under `count` condition in Clevr4 [12]. As shown in this figure, the image features corresponding to numbers 1 to 10 are aligned in a radial direction from the bottom left. This indicates that our method can enhance the rankable property [23] in the original model representation space.

Further qualitative results. In Fig. S7 and Fig. S8, we show additional qualitative results of our method and comparisons for diverse conditions. Given a *query* image and a specific condition, our method reflects the condition and retrieves *database* images accordingly, even distinguishing the differences within the same category (such as `bag category, shoes category`). Moreover, under the `action` condition, our method can disentangle the action attribute with the attributes of the subject (*e.g.*, age, gender), whereas the baseline struggles to retrieve diverse images. These results demonstrate the condition fidelity of our approach under a wide range of conditions.

E. Limitations

While our method achieves substantial gains in retrieval performance, we have a few limitations. First, our work focuses on the core attributes present in the image, which may limit

Table S6. **Quantitative comparison of mean Average Precision (mAP) across VLMs with ViT-L.**

	Stanford40			Fine-grained Image Classification						
Method	Action	Location	Mood	Cat	Dog	Instrument	Flower	Car	Aircraft	Food
CLIP-L	45.3	44.1	54.0	46.7	49.1	35.9	83.6	42.4	29.0	59.0
SigLIP-L	56.5	51.3	56.7	63.1	68.8	53.0	88.9	76.5	60.2	69.7
Ours (CLIP-L)	68.9	55.5	**63.2**	80.1	87.0	70.8	89.7	65.5	43.4	71.5
Ours (SigLIP-L)	**72.2**	**60.0**	62.0	**87.8**	**89.8**	**79.5**	**97.7**	**86.3**	**72.7**	**76.6**

(a) real-world datasets.

	Clevr4				CLAY-Object			CLAY-Human		
Method	Shape	Color	Texture	Count	Color	Category	Subcategory	Age	Action	Background
CLIP-L	71.1	17.1	18.7	13.5	12.9	70.3	48.8	47.3	55.0	39.1
SigLIP-L	83.6	18.7	19.4	13.0	15.3	67.5	60.8	48.1	57.6	54.7
Ours (CLIP-L)	79.5	64.8	24.9	19.9	47.2	**95.1**	78.2	**68.9**	**81.9**	82.3
Ours (SigLIP-L)	**93.1**	**67.6**	**25.6**	**25.8**	**64.9**	94.0	**83.3**	67.6	81.42	**91.6**

(b) synthetic datasets.

Table S7. **Quantitative comparison of Recall@1,2,3 on GeneCIS benchmark.** We evaluate the retrieval performance on the "Focus Attribute" subset, where each query has a single ground-truth match. Recall@1, @2, and @3 are reported as the performance metric. For CIR methods, we report the values provided from MagicLens [28].

Method	Recall@1	Recall@2	Recall@3
CLIP-B	17.8	30.0	40.4
GeneCIS	19.5	31.8	42.2
CIReVL	17.9	29.4	40.4
MagicLens	15.5	28.4	39.1
SEARLE	17.0	29.7	40.7
Ours (CLIP-B)	**24.4**	**38.3**	**50.5**

Table S8. **Quantitative result of multi-conditioned retrieval**

	Stanford40	
	Action + Mood	**Action + Location**
GeneCIS	54.1	47.2
GeneCIS†	65.2	55.5
Ours (CLIP-B)	68.3	57.9
Ours (CLIP-L)	**72.0**	**58.5**

its applicability to the general scenarios, especially when the task involves focusing on multiple objects. Secondly, as our method depends on the feed-forward visual features extracted from VLM, it may not recover attributes that are not encoded in the original representations. For example, prior work [1] shows that the CLIP image encoder favors larger objects, suggesting that the small objects may not be fully

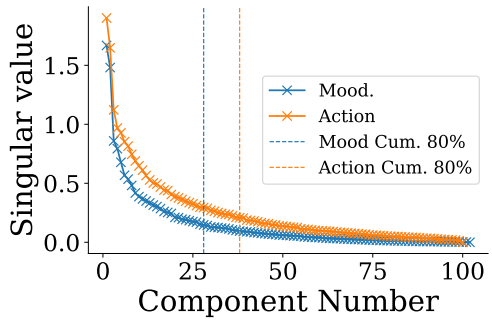


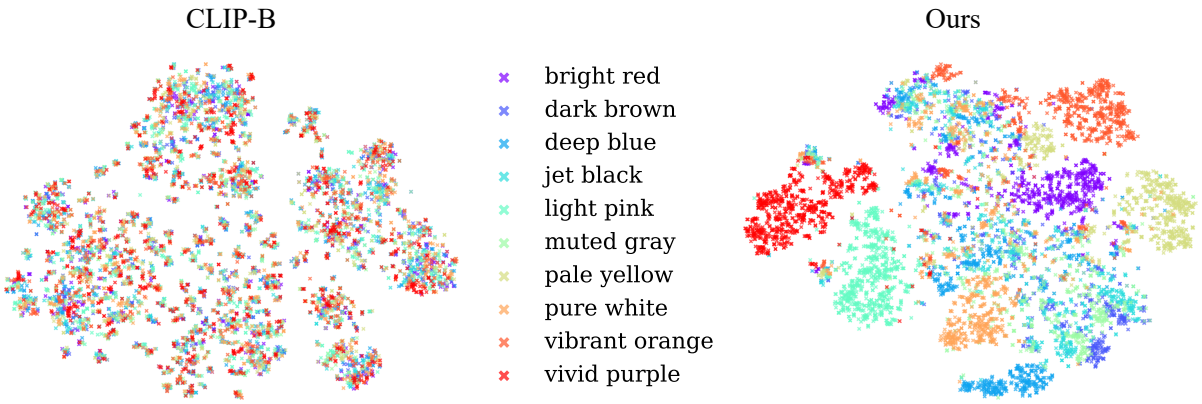
Figure S3. **Scree Plot**

captured and thus our method may fail in this case. However, these limitations can be alleviated by allowing users to specify their conditions in a more direct manner (*e.g.*, segmentation masks), which may help recover the attributes and enable our method to perform retrieval effectively.

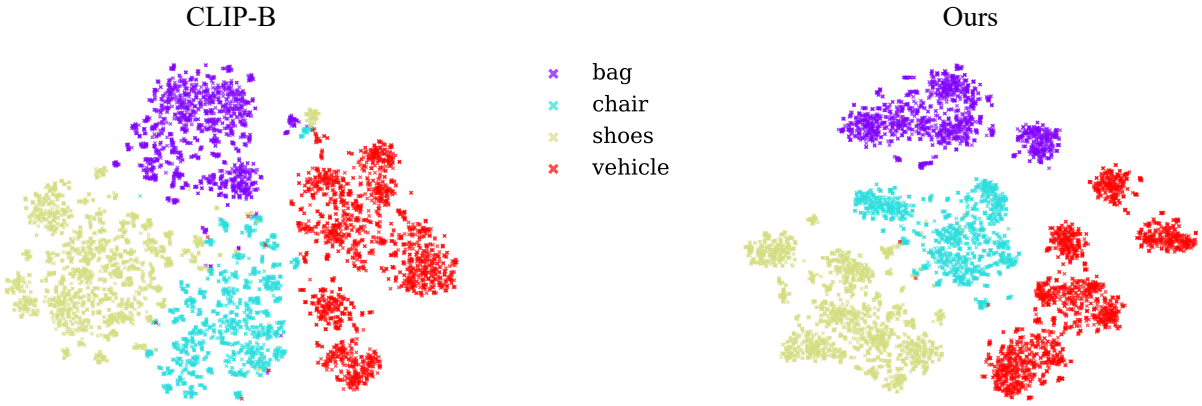
F. Ethical Considerations

Our synthetic dataset CLAY-EVAL involves human images fully generated by image generation model, FLUX [4]. It does not include any real individuals, and we designed both text prompts and generation process to ensure this. The demographic categories referenced—specifically European, Asian, Indian, Middle Eastern, African, and Latino (Hispanic)—are sourced from Fair Diffusion [9] and FairFace [13]. We acknowledge the limitations of these categories, which represent broad, oversimplified groupings and contain inconsistent levels of granularity (e.g., mixing continental, national, and ethnic labels). These labels were used solely for the purpose of ensuring sample diversity in this

paper. Our model is not supplied with and does not make use of this information at any stage of training or inference. To promote safe and responsible use, we performed manual screening step to remove samples containing harmful content. Nevertheless, we recognize the misuse of synthetic human images remains, and we encourage users to utilize the dataset with responsibility and follow the ethical guidelines.

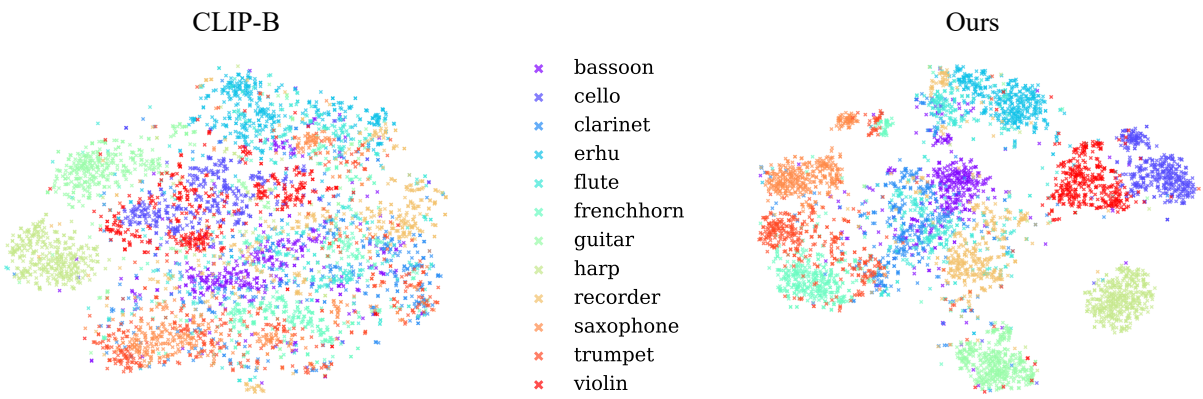


(a) condition c: Color



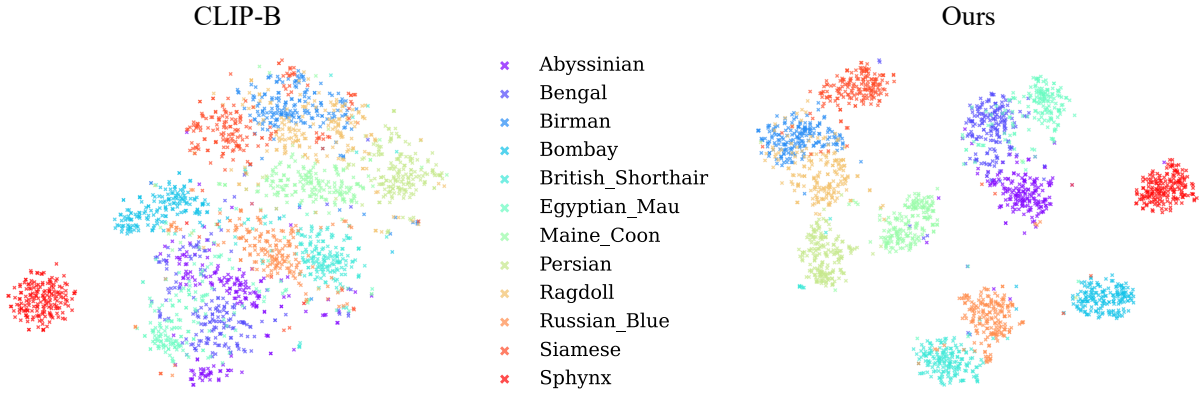
(b) condition c: Category

Figure S4. **T-SNE visualization on CLAY-EVAL-Object dataset.**



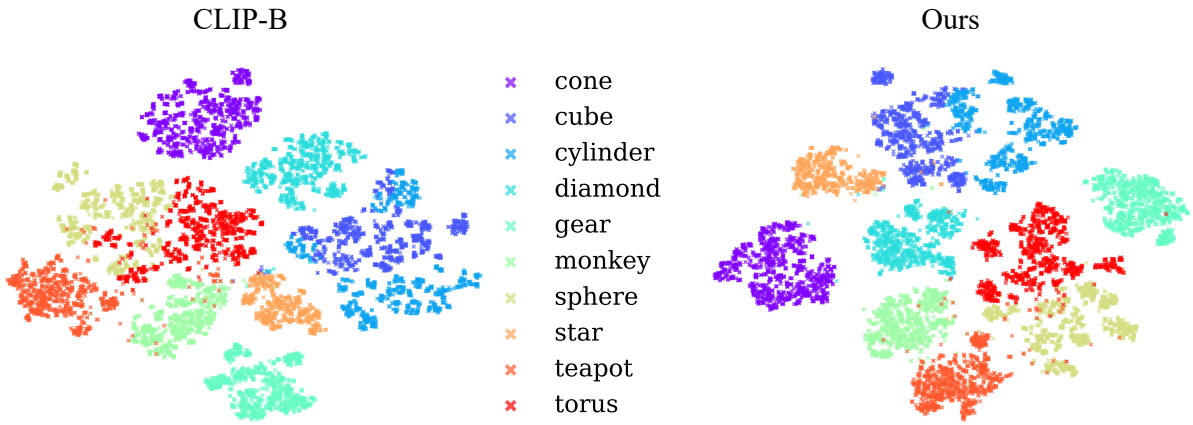
(a) condition c: Instrument

(b) condition c: Dog species

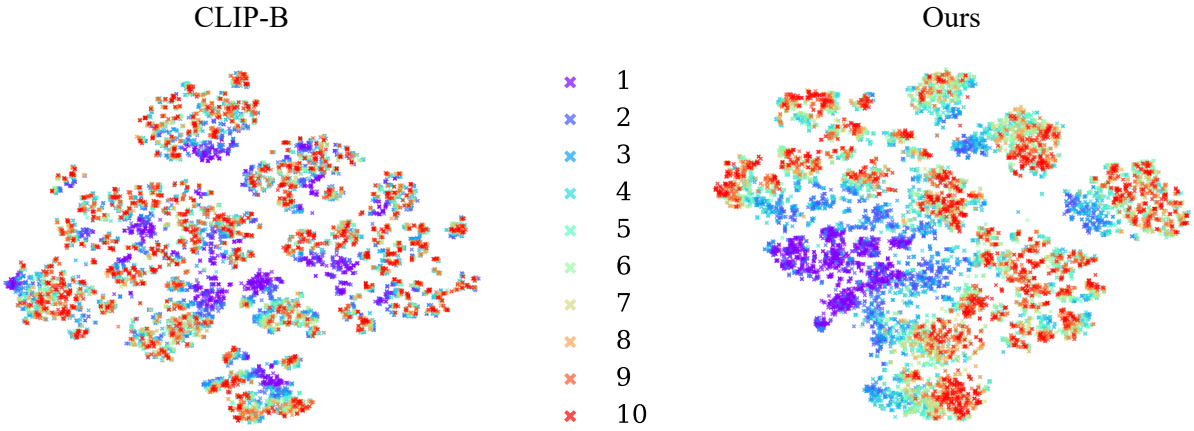


(c) condition c: Cat species

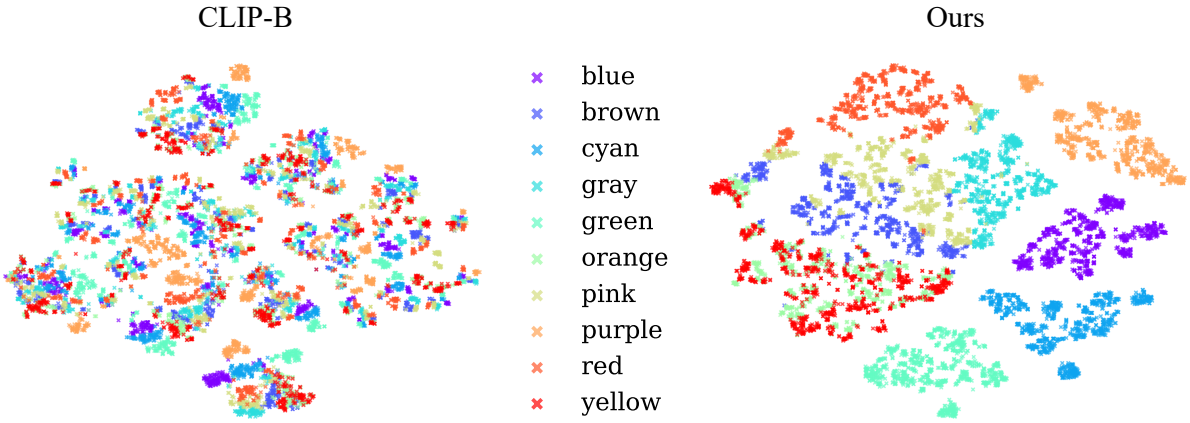
Figure S5. **T-SNE visualization on fine-grained classification datasets.**



(a) condition c: Shape



(b) condition c: Count



(c) condition c: Color

Figure S6. **T-SNE visualization on Clevr4 dataset.**

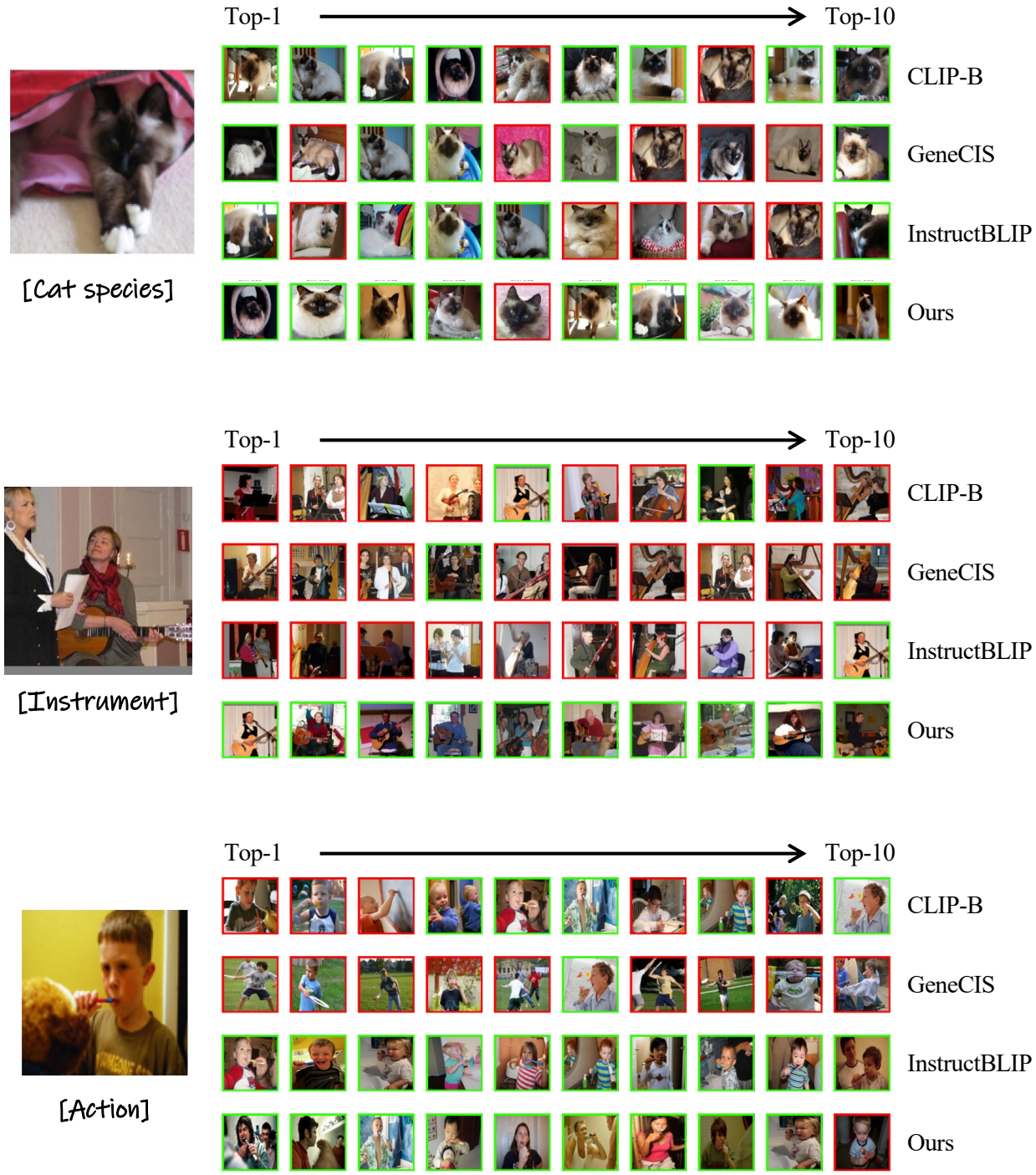


Figure S7. **Qualitative results on real datasets.**

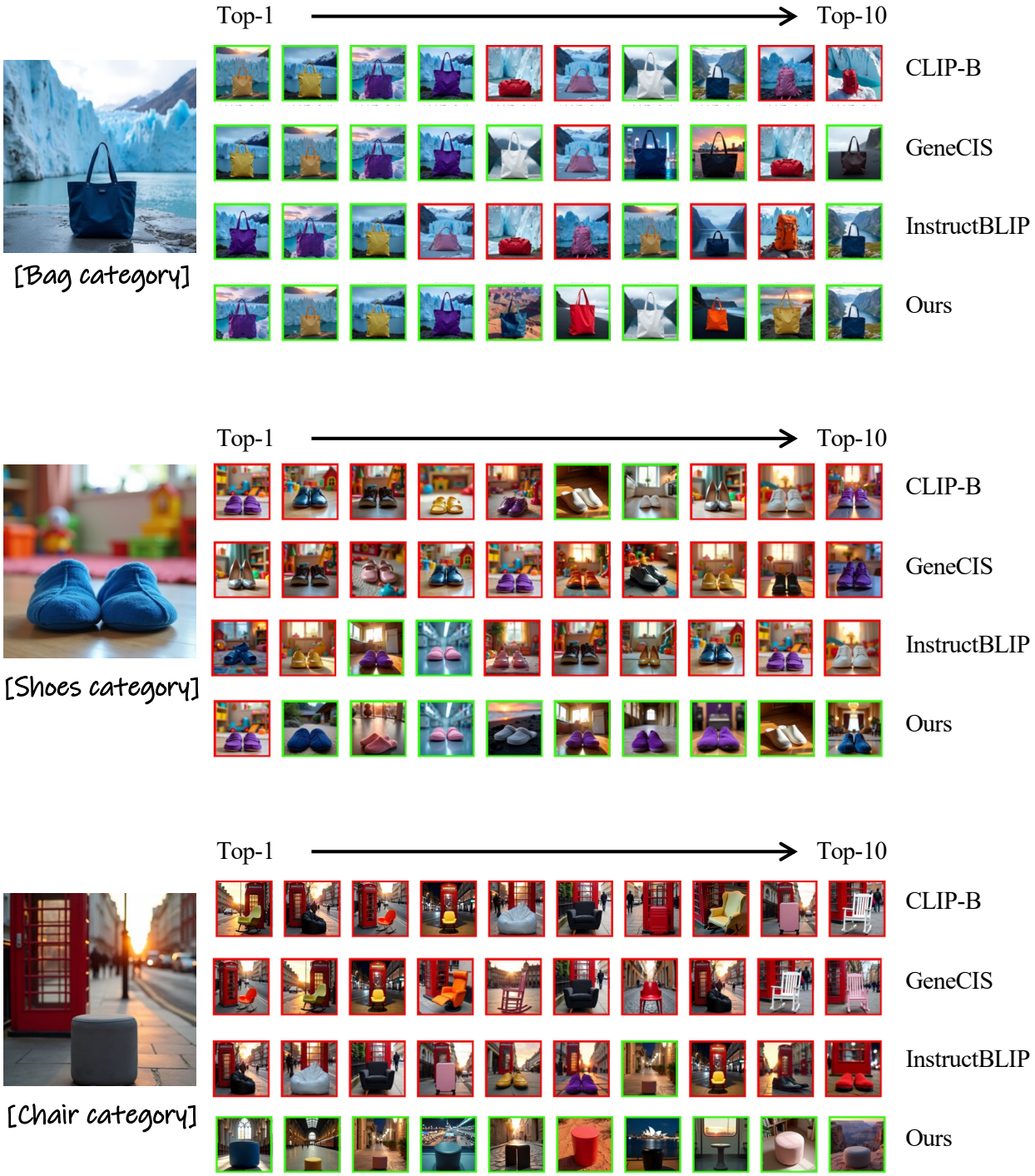


Figure S8. **Qualitative results on synthetic datasets.**

References

[1] Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Analyzing clip's performance limitations in multi-object scenarios: A controlled high-resolution study. *arXiv preprint arXiv:2502.19828*, 2025. 7

[2] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, 2023. 6

[3] Davide Berasi, Matteo Farina, Massimiliano Mancini, Elisa Ricci, and Nicola Strisciuglio. Not only text: Exploring compositionality of visual representations in vision-language models. In *CVPR*, 2025. 1

[4] Black Forest Labs. Flux.1. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. 1-dev. 2, 7

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 5

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 6

[7] Sara Dorfman, Dana Cohen-Bar, Rinon Gal, and Daniel Cohen-Or. Ip-composer: Semantic composition of visual concepts. In *ACM Transactions on Graphics (SIGGRAPH)*, pages 1–11, 2025. 5

[8] P.T. Fletcher, Conglin Lu, S.M. Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8): 995–1005, 2004. 1

[9] Fabian Friedrich, Moritz Brack, Lukas Struppek, David Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 7

[10] Søren Hauberg. Directional statistics with the spherical normal distribution. In *2018 21st international conference on information fusion (FUSION)*, 2018. 1

[11] Cheng-Yu Hsieh, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Chun-Liang Li, Ranjay Krishna, Oncel Tuzel, and Hadi Pouransari. Focallens: Instruction tuning enables zero-shot conditional image representations. *arXiv preprint arXiv:2504.08368*, 2025. 6

[12] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 6

[13] K. Karkkainen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, 2021. 7

[14] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *ICLR*, 2024. 6

[15] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 5

[16] Mingxin Li, Yanzhao Zhang, Dingkun Long, Chen Keqin, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026. 6

[17] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022. 1

[18] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[19] Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025. 6

[20] Quang-Binh Nguyen, Minh Luu, Quang Nguyen, Anh Tran, and Khoi Nguyen. Csd-var: Content-style decomposition in visual autoregressive models. In *ICCV*, 2025. 5

[21] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 5

[22] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012. 5

[23] Ankit Sonthalia, Arnas Uselis, and Seong Joon Oh. On the rankability of visual embeddings. In *NeurIPS*, 2025. 6

[24] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *CVPR*, 2023. 6

[25] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 5

[26] Bangpeng Yao and Li Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 5

[27] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 5

[28] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: self-supervised image retrieval with open-ended instructions. In *ICML*, 2024. 6, 7