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

Sohwi Lim Lee Hyoseok Jungjoon Park Tae-Hyun Oh

KAIST

Figure 1. Our proposed concept-based conditional image retrieval method retrieves images focusing on the semantic aspects specified by the text condition. Given a query image, our method adaptively computes conditioned similarity by modulating the similarity space to align with various conditions, *e.g.*, `species`, `location`, `action`, `category`, and `color`.

Abstract

*Human perception of visual similarity is inherently adaptive and subjective, depending on the users' interests and focus. However, most image retrieval systems fail to reflect this flexibility, relying on a fixed, monolithic metric that cannot incorporate multiple conditions simultaneously. To address this, we propose **CLAY**, an adaptive similarity computation method that reframes the embedding space of pretrained Vision-Language Models (VLMs) as a text-conditional similarity space without additional training. This design separates the textual conditioning process and visual feature extraction, allowing highly efficient and multi-conditioned retrieval with fixed visual embeddings. We also construct a synthetic evaluation dataset **CLAY-EVAL**, for comprehensive assessment under diverse conditioned retrieval settings. Experiments on standard datasets and our proposed dataset show that CLAY achieves state-of-the-art retrieval accuracy and notable computational efficiency compared to previous works. Our code and datasets will be publicly released upon acceptance.*

1. Introduction

In the era of unprecedented data scale, humans seek to efficiently and accurately identify the information of interest in the overwhelming data flow. Within this context, retrieval serves as a computational mechanism that enables us to find what truly matters and what we need. In the computer vision area, despite the remarkable advances in large-scale image retrieval tasks, most approaches still rely on static definitions of visual similarity [6, 12, 35, 47]. By contrast, humans perceive visual similarity in a flexible and adaptive manner, selectively focusing on different aspects of an image depending on a user's interest. For example, one may seek the same object itself, while another may want to see the overall mood of the image. This underscores the need for conditional retrieval systems that can reflect various human

Table 1. Our contextual conditional similarity computation method provides high retrieval accuracy while maintaining efficiency under diverse conditions. We further support a multi-conditioned retrieval scheme, whereas previous works [17, 39] do not. † denotes a modified version of the GeneCIS, where database features are additionally incorporated with conditional text.

	GeneCIS	GeneCIS†	FocalLens	CLAY
Training-free	✗	✗	✗	✓
Retrieval accuracy	✗	✓	✓	✓
Dynamic efficiency	✓	✗	✗	✓
Multi-condition	✗	✗	✗	✓

attention-related visual cues.

In response to this demand, prior studies have explored two main directions depending on which aspects of the query image are focused or modified during retrieval: one focuses on particular attributes within the query image [17, 39, 40], while the other aims to change specific visual contexts of query image to match the target images [1, 14, 18, 20, 37, 48]. Although the latter direction has been actively studied, the former direction, which seeks to capture specific attributes in a dynamic manner, remains relatively underexplored despite its importance. Most studies have addressed the latter direction in a training-based manner, where models are trained to perform retrieval under given conditions by learning condition-specific representations [17, 39]. However, these methods require not only computational resources for the training stage but also paired query-target image data for each condition [39], which limits their working regime to closed-set conditions. Moreover, to maintain retrieval accuracy, these trained conditional feature extractors incur computational overhead at inference time because all features of database images need to be recomputed through the model from scratch, whenever the user condition changes [17, 40]. These limitations hinder the applicability and capacity of conditional retrieval systems in diverse, real-world scenarios.

To overcome these challenges, we propose CLAY, a training-free and adaptive conditional similarity computation method that transforms the similarity computation space of pretrained Vision-Language Models (VLMs) [35, 47], according to users' interest. By decoupling the conditioning procedure from visual feature extraction, our approach contextually modulates the similarity computation space of VLMs into a conditional similarity space. This eliminates the need to completely recompute embeddings of database images under varying user conditions by keeping the original visual embeddings fixed. To achieve this, we construct a conditional similarity space within the VLMs' representation space that respects its underlying non-Euclidean geometry, and define its textual concept subspace leveraging a subspace–construction procedure similar to [9, 29]. Building on this, we further demonstrate that our method is easily extensible to multi-conditional retrieval scenarios, unlike previous methods dedicated to single-condition settings [17, 39]. Due to a lack of a standard benchmark of multi-conditional retrieval task, we construct a synthetic evaluation dataset containing diverse human and object images and various conceptual condition pairs, allowing evaluations under multi-condition scenarios. Evaluation on diverse real and synthetic datasets shows that our method achieves strong retrieval accuracy and notable computational efficiency, pushing the Pareto-front of the trade-off between performance and efficiency in practice. We summarize our contributions as:

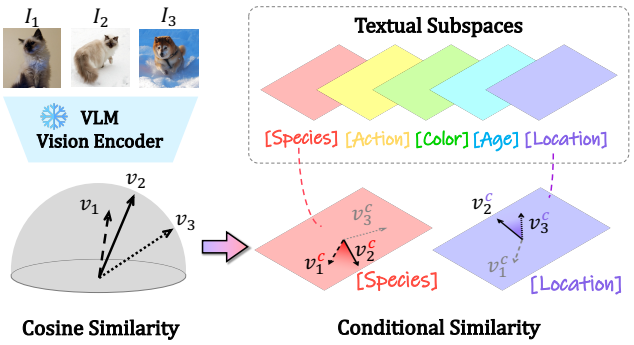


Figure 2. **Illustration of the concept of CLAY.** Our method adaptively computes conditional similarity between images by modulating the original similarity space into a conditional similarity space within the representation space of VLMs.

- We propose an efficient, training-free, state-of-the-art conditional visual similarity computation method that can contextually adapt to various conditions without recomputing database features at inference.
- Our method supports multi-conditioned image retrieval scenarios, enabling flexible retrieval settings beyond single-condition settings.
- We construct a synthetic evaluation dataset that contains diverse humans and objects with conceptual condition pairs to facilitate the evaluation under diverse conditional retrieval scenarios.

2. Related Work

Image-to-image retrieval is a practical and fundamental problem in the computer vision area [16]. Conventional methods on image-to-image retrieval mainly focused on measuring how visually similar two images are, relying on handcrafted local descriptors [8, 27]. With the emergence of deep learning, subsequent studies leveraged Convolutional Neural Networks [5, 31, 32, 43] to extract higher-level semantic features beyond low-level visual cues. However, relying solely on static visual feature similarity from such models can be insufficient, as human preferences may differ on multiple contextual factors. This limitation suggests that the retrieval systems need to be designed to consider further conditions, *i.e.*, conditional image retrieval. In this work, we explore

conditional image retrieval, which considers adaptive alignment with user intentions.

Vision-Language Models (VLMs) are trained to embed image and text modalities into a shared embedding space [35, 47]. They also provide semantically well-aligned visual representations by leveraging language as a semantic reference [42]. Recent studies [9, 29] have further explored the structure of the VLM embedding space to enable text-guided control over visual embeddings for personalized image generation and editing. For instance, Dorfman et al. [9] project visual embedding into a text-based subspace, thereby extracting the text-related visual contexts useful for compositional image generation. However, these approaches primarily focus on aligning visual representations with textual semantics, rather than modeling their relative relationships. We further develop this idea to effectively capture the relationships among visual features within the textual subspace, taking into account the hyperspherical nature of the embedding manifold for accurate relationship modeling.

Conditional image retrieval. With the increasing importance of retrieving images that satisfy user-specified conditions, prior studies have proposed utilizing text conditions and incorporating visual features to focus on specific attributes [17] in the image or to modify particular contexts [1, 26, 44, 48]. As an early work, Conditional Similarity Networks [40] alleviated the limitation of the single similarity metric of embedding methods by employing conditioning masks to select condition-specific embedding dimensions. To specify this task, GeneCIS [39] introduced a benchmark that categorizes text conditions into two types: *focus on* and *change*, and proposed a training-based method that fine-tunes the image encoder and condition feature modulator from the paired dataset. Following this, most subsequent studies have tackled the *change* type, retrieving images that differ only in specific attributes from the query image, known as Composed Image Retrieval (CIR) [1, 14, 18, 21, 26, 44, 48]. Another line of work [17] aims to retrieve images by focusing on specific attributes within a query image, leveraging VLMs or multimodal Large Language Models (mLLMs). We explore the latter direction, which is relatively underexplored compared to CIR, and leverage the joint embedding space of VLMs to effectively perform conditional image retrieval.

3. CLAY: Conditional Similarity Modulation

We propose a conditional visual similarity computation method leveraging VLMs [35, 47] that consider hyperspherical manifold. We begin by describing the problem definition of conditional retrieval in Sec. 3.1, followed by the formulation of our conditional similarity in Sec. 3.2.

3.1. Problem Definition

Previous conditional retrieval methods [17, 39] typically compute condition-based visual similarity by modifying image representations with text conditions through a feature modulator. In contrast, we reformulate the visual similarity space itself, allowing conditional retrieval with fixed embeddings. We first categorize the conditional similarity computation methods into two formulations: symmetric and asymmetric methods based on how the text condition c is incorporated into the image representations. Given *query* image I_q and *database* images I_d, the conditional similarity between I_q and I_d can be computed following these two formulations:

$$\texttt{csim}_{\text{sym}}(I_q, I_d \mid c) = d\big(m(I_q, c),\, m(I_d, c)\big), \quad (1)$$

$$\texttt{csim}_{\text{asym}}(I_q, I_d \mid c) = d\big(m(I_q, c),\, I_d\big), \quad (2)$$

where $d(\cdot, \cdot)$ denotes a similarity function such as cosine similarity. The $m(I, c)$ represents a *modulator* that integrates input image I and the given condition c.

Symmetric vs. asymmetric. Compared to the symmetric case (Eq. 1), the asymmetric computation (Eq. 2) does not forward the *database* images through the modulator m with the text condition. A representative method GeneCIS [39] follows the asymmetric formulation, where only *query* features are conditioned on the text input. However, despite the relative efficiency of this asymmetric formulation, the *database* features will remain independent of the given condition. Consequently, the retrieved results rely on the condition-agnostic representations, potentially leading to suboptimal retrieval performance when guiding the search. To further support this, we experimentally compare the performance by applying both the symmetric and asymmetric formulations to GeneCIS, in Sec 5.

Design of modulator m. Existing methods [17, 39] design the modulator m through the neural network that incorporates text and image inputs. In other words, the modulator is coupled with the visual feature extraction and the conditioning process through a neural network coupler to obtain conditional visual features. This design requires a full forward pass through the network for each condition to obtain the corresponding conditional features. In the symmetric form, it may lead to computational overhead and limit its practicality for adaptive conditional retrieval. In contrast, we decouple the conditioning process from the visual feature extraction within the modulator m, and propose a conditioning process that does not rely on the input visual feature. We achieve this by proposing the conditional similarity space modulation scheme, which directly leverages the visual features from a pretrained VLM and projects them into a conditional similarity space. This decoupling enables efficient adaptive retrieval in the symmetric form by eliminating the need to re-encode the database features for varying conditions.

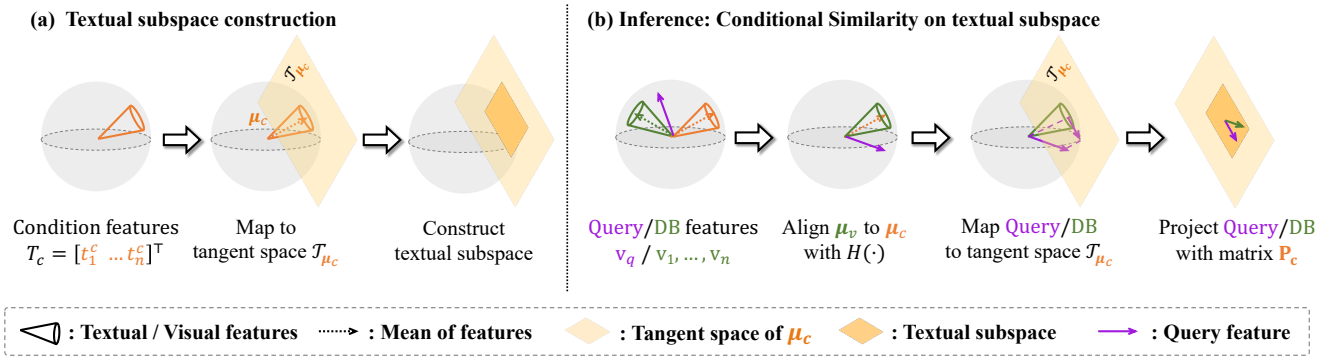


Figure 3. **Conditional similarity computation pipeline.** (a) Given a condition, we construct the manifold-aware textual subspace with the condition text features in advance, and generate the condition-aware projection matrix $\mathbf{P}_c$. (b) At inference, we compute the conditional similarity between the query and database images by projecting the visual features onto the textual subspace with $\mathbf{P}_c$.

3.2. Visual Similarity Modulation

As mentioned, since we utilize text modality for describing the condition, we adopt pretrained VLMs to leverage their joint embedding space. VLMs typically consist of a vision encoder f_I and text encoder f_T. For each *query* I_q, *database* image I_d and condition c, we define $\mathbf{v}_q = f_I(I_q)$, $\mathbf{v}_d = f_I(I_d)$, and $\mathbf{t} = f_T(c)$, where f_I and f_T denote the vision and text encoder, respectively.

The main idea of conditional similarity space modulation is to project visual features onto the textual subspace, which captures the condition-aware relationships among visual features. To achieve this, we derive a projection matrix $\mathbf{P}_c$ by performing singular value decomposition (SVD) on the text condition feature matrix. Specifically, following previous works [9, 29], we first generate condition-related textual prompts through Large-Language Model (LLM) with the template: `a photo of` {c}. The generated text prompts are then encoded through a text encoder f_T and their text embeddings $\mathbf{t}$ are concatenated to form the text feature matrix $\mathbf{T}_c = [\mathbf{t}_1^c, \dots, \mathbf{t}_n^c]^\top \in \mathbb{R}^{n \times d}$. Previous works directly apply SVD on these text features to generate the projection matrix; however, this operation assumes Euclidean structure and therefore ignores the intrinsic geometry of the embedding manifold. In contrast, since VLM embeddings lie on a unit hypersphere, this Euclidean assumption fails to capture their underlying geometry.

Manifold-aware textual subspace construction. To address this, we take into account the hyperspherical nature of the embedding manifold in VLMs to accurately model the relationship and derive a manifold-aware projection matrix $\mathbf{P}_c$. We construct the textual subspace as locally geodesic submanifold rather than linear subspace, to alleviate the distortion from Euclidean projection similar to previous works [2, 11]. Under mild curvature assumptions, this submanifold can be locally approximated by its tangent space at a reference point. Previous works [10, 25] observe that the embedding geometry of each modality tends to be concentrated within a conical region, indicating that most embeddings lie closer to a common mean direction. Hence, we leverage the validity of the local tangent space approximation under this geometric property. Formally, any point $\mathbf{x} \in \mathcal{S}^{d-1} \setminus \{-\boldsymbol{\mu}\}$ can be mapped onto the tangent space $\mathcal{T}_{\boldsymbol{\mu}} = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}^\top \boldsymbol{\mu} = 0\}$ via *logarithm map* [15]:

$$\log_{\boldsymbol{\mu}}(\mathbf{x}) := (\mathbf{x} - \boldsymbol{\mu}(\mathbf{x}^\top \boldsymbol{\mu}))\frac{\theta}{\sin(\theta)}, \quad \theta = \arccos(\mathbf{x}^\top \boldsymbol{\mu}). \tag{3}$$

We consider the normalized mean $\boldsymbol{\mu}_c$ of text features $\mathbf{T}_c$ as a reference point for defining the tangent space, and map text features from the unit hypersphere manifold to the tangent space of $\boldsymbol{\mu}_c$ with *logarithm* map. Then, we apply SVD on these mapped text features:

$$\log_{\boldsymbol{\mu}_c}(\mathbf{T}_c) = \left[\log_{\boldsymbol{\mu}_c}(\mathbf{t}_1^c) \cdots \log_{\boldsymbol{\mu}_c}(\mathbf{t}_n^c)\right]^\top, \quad \log_{\boldsymbol{\mu}_c}(\mathbf{T}_c) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top. \tag{4}$$

We utilize the top-k right singular vectors to construct textual subspace, defined as $\text{span}(\mathbf{V}_k)$, and obtain the projection matrix $\mathbf{P}_c = \mathbf{V}_k \mathbf{V}_k^\top$. By precomputing these textual subspaces for each condition (*i.e.*, projection matrices), we only need to modulate the conditional similarity space using precomputed projection matrices at inference, accordingly. Specifically, we project the query and database visual features from pre-trained VLMs into the space that satisfies the desired conditional relationships, without re-encoding.

Inference. Now we have a projection matrix $\mathbf{P}_c$ which projects features onto the textual subspace within the tangent space of $\boldsymbol{\mu}_c$. Our ultimate goal is to model the intra-relationship among visual features within this condition-aware textual subspace. As mentioned above, the *logarithm map* onto the tangent space is valid under mild curvature assumptions. However, due to the conic effect [10, 25], naive projection of visual features cannot maintain this property.

To mitigate this, we apply rotation $H(\cdot)$ to align the mean of the database visual features $\boldsymbol{\mu}_{v_d}$ with the mean of text features $\boldsymbol{\mu}_c$, without altering the intra-relationship among the visual features. We then apply the logarithm map to map the rotated visual features $H(\mathbf{v})$ onto the tangent space at $\boldsymbol{\mu}_c$, and subsequently project them onto the textual subspace with projection matrix $\mathbf{P}_c$. Finally, similar to standard similarity computation, we compute the similarity between the query feature $\mathbf{v}_q$ and the database features $\mathbf{v}_d$ by using cosine similarity. The proposed conditional similarity between the query image I_q and the database image I_d under the condition c is as follows:

$$\begin{aligned} m_{\text{CLAY}}(\mathbf{v}, c) &:= \mathbf{P}_c \log_{\boldsymbol{\mu}_c}(H(\mathbf{v})), \\ \texttt{csim}_{\text{CLAY}}(I_q, I_d \mid c) &:= d\big(m(I_q, c), m(I_d, c)\big), \end{aligned} \tag{5}$$

where $d(\cdot, \cdot)$ corresponds to the cosine similarity. The detailed pipeline is provided in the supplementary material.

4. Our CLAY-EVAL Dataset

We provide a novel synthetic evaluation dataset CLAY-EVAL for conditional image retrieval. To evaluate our context-aware conditional image retrieval setting, each image in the dataset should be annotated with multiple labels, allowing flexible clustering under different conditions. While several datasets are classified with particular conditions, their scalability is limited because real-world images were manually annotated [23], they only have a single condition [30, 33, 41, 45, 46], or they include only simple simulated 3D geometric objects (*e.g.*, cone, sphere, and cube) with a simple background [19]. To address this, we further build a synthetic dataset using a powerful open-source diffusion model [3], whose reliability allows for the controlled generation of high-quality and diverse image samples.

To construct an evaluation dataset aligned with our objectives, we establish three key design principles: disentanglement, compositionality, and naturalness, inspired by Li et al. [24]. Based on these, we structure our dataset into two main entities, *object* and *human*. Each dataset consists of core attributes, which serve as the primary query conditions, and diversity attributes, which control for visual variance and bias. We define three core attributes for each entity, namely category, sub-category, and color for the *object* entity, and age, action and background for the *human* entity.

We generate images with structured prompts derived from all attribute combinations. For stringent quality control, we first exclude contradictory combinations via schema-level filtering, and subsequently manually remove generated images exhibiting low text-visual alignment.

As shown in Fig. 4, our synthetic dataset covers a wide range of instances and paired images, enabling comprehensive context-aware conditional retrieval evaluation. Our final synthetic dataset consists of 7,325 *object* images and 6,745 *human* images, and detailed lists of all attribute instances and curation process are provided in the supplementary material.

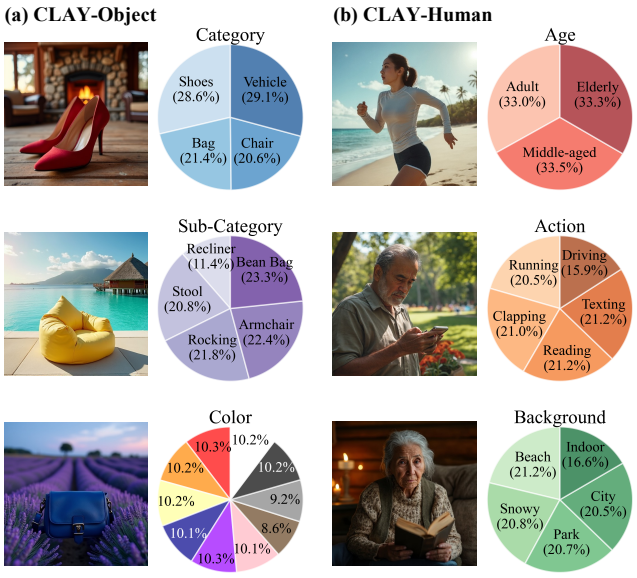


Figure 4. **Our CLAY-EVAL dataset statistics.** We construct a synthetic dataset with diverse condition annotations, consisting of (a) *object* entity and (b) *human* entity. For both, the left column shows sample images demonstrating visual naturalness, and the right column visualizes the distributions of key attributes, showing diversity. Percentages are truncated to one decimal place and annotation text labels are abbreviated. See supplementary material for full details.

5. Experiments

5.1. Experimental Setup

Evaluation datasets and metric. We evaluate the conditional retrieval performance on a wide range of datasets, including both real-world datasets and our synthetic dataset. For real-world datasets, we utilize fine-grained image classification datasets [4, 22, 28, 30, 33, 41, 45], and further conduct on Stanford40 [46] with the human annotated labels from [23]. In these fine-grained classification datasets, we consider each category as condition, for example, in Stanford40, the condition we supposed is `action`. For the synthetic evaluation, we employ CLEVR4 [19] with the conditions `shape`, `color`, `texture`, and `count`, as well as our generated synthetic dataset. We split each dataset into *query* and *database* with 1:9 ratio, and in the evaluation with `subcategory` condition, we report the averaged performance from each categorized database. For the evaluation metric, we adopt the standard metric mean Average Precision (mAP) following conventional image retrieval tasks [13, 34, 36], unless otherwise noted.

Competing methods. To the best of our knowledge, GeneCIS [39] is the only work that is closely related to our problem. As mentioned above, they follow the asymmet-

Table 2. **Quantitative comparison of mean Average Precision (mAP) on single conditional datasets.** We evaluate single conditional retrieval performance across various datasets, where each condition corresponds to a specific attribute (*e.g.*, `action`, `cat species`, `flower type`). (a) Real-world datasets and (b) synthetic datasets are presented separately. We compare the baseline Vision-Language Models (VLMs) and competing methods that support conditional retrieval. For a fair comparison, we additionally report GeneCIS†, the extended results of GeneCIS using a symmetric conditional similarity computation, whereas the original GeneCIS adopts an asymmetric formulation. The **best** and **second-best** results are highlighted.

Method	Stanford40			Fine-grained Image Classification						
	Action	Location	Mood	Cat	Dog	Instrument	Flower	Car	Aircraft	Food
CLIP-B	43.0	47.0	53.0	37.5	37.9	27.1	70.1	30.5	20.9	47.4
SigLIP-B	54.8	52.7	56.4	55.9	59.3	40.6	86.6	64.6	46.7	61.3
GeneCIS	50.0	50.9	51.8	24.7	24.7	24.7	58.5	16.9	17.2	44.8
GeneCIS†	63.5	52.3	57.3	32.7	33.9	42.0	72.3	29.8	23.5	50.8
InstructBLIP	63.1	54.4	60.3	50.6	56.8	37.1	76.5	27.6	15.7	50.6
Ours (CLIP-B)	66.0	55.4	57.9	72.7	79.4	58.0	80.4	51.2	28.4	58.8
Ours (SigLIP-B)	66.2	59.5	58.7	82.1	84.7	63.4	92.7	78.0	57.9	66.1

(a) real-world datasets.

Method	Clever4				CLAY-Object			CLAY-Human		
	Shape	Color	Texture	Count	Color	Category	Subcategory	Age	Action	Background
CLIP-B	61.7	19.9	18.1	13.9	12.9	68.9	42.4	50.4	54.4	41.8
SigLIP-B	77.7	20.0	18.6	13.0	14.8	69.6	62.0	47.7	56.7	53.6
GeneCIS	47.9	16.6	16.0	11.5	12.5	80.1	38.4	45.1	62.4	46.2
GeneCIS†	72.6	17.4	18.1	12.7	11.9	84.3	39.6	44.1	71.3	48.9
InstructBLIP	83.4	27.4	22.8	17.9	14.4	75.6	61.4	45.0	72.2	73.7
Ours (CLIP-B)	72.1	67.9	22.2	21.8	47.8	93.4	65.8	71.3	81.3	75.5
Ours (SigLIP-B)	88.6	73.2	26.1	22.7	63.3	94.3	81.9	61.0	80.4	81.9

(b) synthetic datasets.

Table 3. **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 [48].

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 4. **Quantitative comparison of mAP on multi-conditional datasets CLAY.** We provide the condition combinations such as `color and category`. We report mAP as an evaluation metric. The best results are highlighted.

Method	CLAY-Object		CLAY-Human			
	Color Category	Color Subcategory	Age Action	Age Background	Action Background	All
CLIP-B	11.6	10.9	35.9	27.1	37.0	32.0
SigLIP-B	13.9	19.7	34.5	32.1	50.5	39.5
InstructBLIP	12.5	14.7	38.6	31.2	74.7	44.1
Ours (CLIP-B)	35.9	38.2	**57.0**	**56.6**	78.6	**58.0**
Ours (SigLIP-B)	**44.7**	**55.0**	49.0	55.7	**81.5**	52.0

ric formulation, only conditioning the *query* features. We further apply GeneCIS in symmetric formulation by also feed-forwarding the *database* features with the conditions through the modulator, noted as GeneCIS†. In addition, we also include InstructBLIP [7] as our competing method, which utilizes CLIP vision encoder with LLM to perform instruction-following tasks. To extract the conditioned visual features in InstructBLIP, we average Q-former output tokens, following FocalLens [17]. The detailed instructions for each condition and implementation details are provided in the supplementary materials.

Implementation details. In our method, we utilize two VLMs, CLIP [35] and SigLIP [47]. To generate condition-related textual prompts, we utilize ChatGPT-5. We truncate the top-k singular vectors and select $k = 50$ for all experiments. Additional details are in the supplementary materials.

5.2. Comparison of Retrieval Accuracy

Table 2 presents the single conditional retrieval accuracy of our method and competing methods across various real

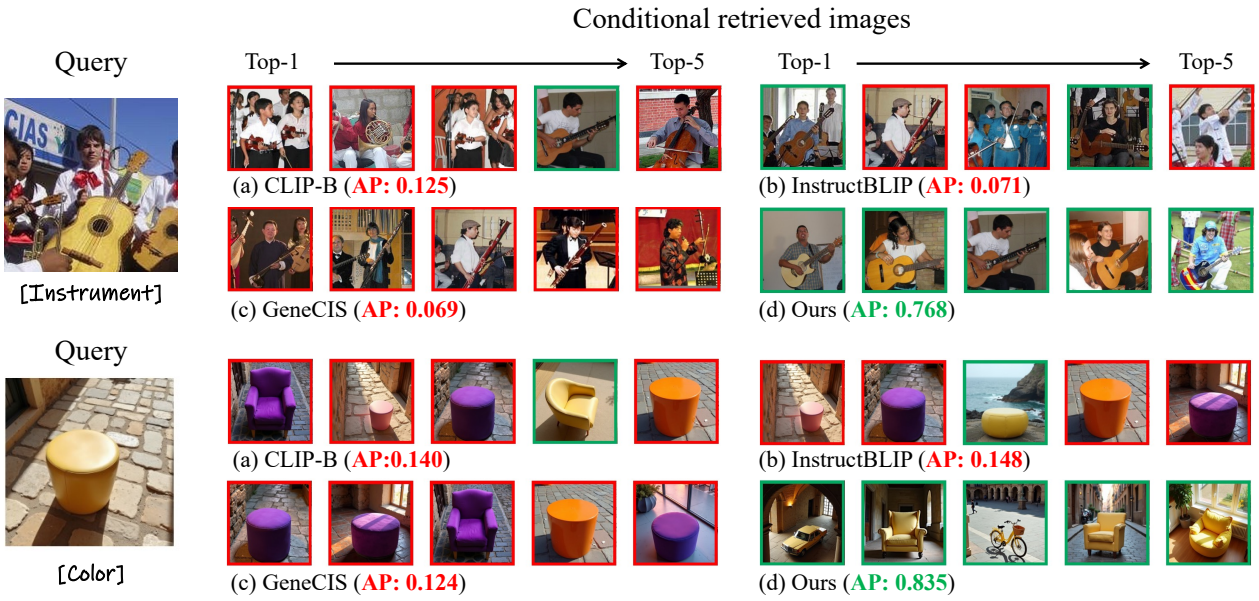


Figure 5. **Qualitative comparison of our method with competing methods.** For each query image and condition text pair, we compare the top-5 retrieved results from (a) CLIP-B, (b) InstructBLIP, (c) GeneCIS, and (d) our method. We also report Average Precision (AP) in each result. Green boxes indicate correctly retrieved images, while incorrect retrievals are shown in red boxes.

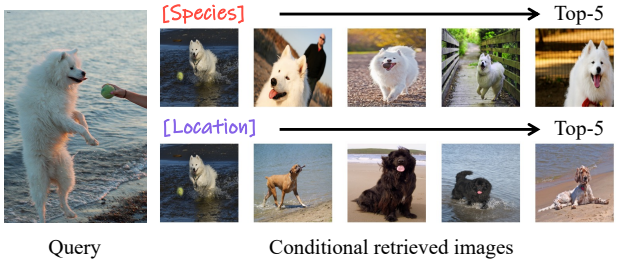


Figure 6. **Qualitative result on Oxfordpets dataset.** We visualize the top-5 retrieved results with condition `dog species` and `location`. Since no ground-truth location labels are available, we present qualitative examples only.

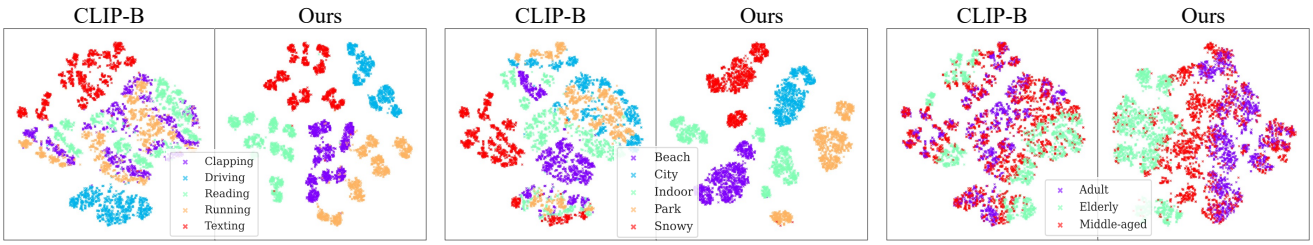
and synthetic evaluation datasets. For clarity, we also report the baseline performance of CLIP-B and SigLIP-B, which cannot input text conditions. As shown in this table, CLAY consistently outperforms competing methods, exhibiting the strong adaptability across diverse datasets and varying conditions, highlighting that similarity modulation can yield highly effective results. GeneCIS, a representative training-based method, performs poorly in the asymmetric setting; as discussed in Sec. 3.1, its symmetric variant (*i.e.*, GeneCIS†) generally achieves better performance. Although GeneCIS shows reasonable quantitative results in the symmetric case, it remains impractical due to its computational overhead. Moreover, while it performs well on several cases (Stanford40), its performance degrades under `dog species` and `car model` conditions compared to the baseline. This implies that while GeneCIS benefits from specific types of text conditions, it has limited generalization capability.

In contrast, our method inherits the strong zero-shot capability of pre-trained VLMs, achieving state-of-the-art performance across diverse datasets and varying condition types without incurring highly computational overhead in the symmetric setting. We further present the quantitative results on GeneCIS benchmark "Focus Attribute" subset, compared to Composed Image Retrieval methods [1, 21, 48]. Table 3 shows CLAY still achieves the best performance when utilizing the same or even smaller backbone (*i.e.*, ViT-B). In Fig. 5, we visualize conditional retrieved results, and it demonstrates that our method achieves better condition-aligned retrieval than the comparison methods.

In addition, CLAY can easily be extended to multi-conditional retrieval by constructing text feature matrix from multiple condition-related textual prompts. Table 4 indicates CLAY performs reliably under the multi-conditional retrieval setting, achieving high accuracy on CLAY-Object and CLAY-Human datasets. Furthermore, as shown in Fig. 6, CLAY effectively reflects diverse conditions within the same query. Figure 7 also shows how the representation space varies under different conditions. This wide range of cover-

Table 5. **Effectiveness of manifold modeling.** We ablate the underlying manifold modeling across datasets. To better capture general trends, we aggregate the results from datasets, and report the averaged mAP. Due to space limitations, we abbreviate manifold modeling as manifold, and CLAY-Human and CLAY-Object as Human and Object, respectively.

	Stanford40	Fine-grained.	Clever4	Object	Human
w/o manifold	57.9	60.5	43.8	65.4	76.8
w/ manifold	**59.8**	**61.3**	**46.0**	**66.9**	**78.3**



(a) condition c: Action (b) condition c: Background (c) condition c: Age

Figure 7. **Representation space visualization with t-SNE.** We report t-SNE of CLIP-B and ours (CLIP-B) on CLAY-Human under condition (a) `action`, (b) `background`, and (c) `age`. The features with the same label are shown in the same color for easy interpretation. Compared to the fixed representation space in CLIP-B, our method forms more discriminative spaces compliant with given conditions.

age is crucial in real-world scenarios, where users may have various intentions that need to be satisfied.

Table 5 shows the performance gains achieved by considering the geometry of VLM's embedding space to preserve relative relationships. Accounting for the hyperspherical nature of the representation space further reduces distortion during space modulation, leading to consistent improvements in retrieval accuracy and highlighting the advantage of considering the underlying geometry of VLM representations.

5.3. Comparison of Inference Time

As mentioned above, a symmetric formulation can be beneficial for improving retrieval accuracy by capturing richer condition-aware visual representations from query and database images. However, coupling the visual feature extraction process with conditioning module introduces non-negligible computational demands, since condition-aware visual representations need to be re-encoded whenever the condition changes or a new one is introduced. In contrast, our method decouples the conditional feature extraction with visual feature extraction and conditioning process, eliminating the need for re-encoding during inference. Figure 8 demonstrates the advantage of CLAY regarding the computing time as additional conditions are given. In this experiment, we sample 100 database images from CLAY-Object and report the averaged inference time of 10 retrievals for a single query on CPU. As the number of conditions increased, the inference time of GeneCIS† increases accordingly since it needs to extract the *database* features for each condition again.

5.4. Analysis of Representation Space

To confirm the effectiveness of similarity space modulation, we visualize t-SNE results comparing with CLIP-B and Ours (CLIP-B) on our CLAY-Human dataset. As shown in Fig 7, the visual representation space is distinctly separated with each `action` and `background` condition, compared to the baseline. This result shows how our method adaptively modulates the original visual features onto the condition-aware subspace, achieving high conditional retrieval performance. Interestingly, applying our method shows a rankable property along the condition axis (*i.e.*, `age`) in Fig 7 (c).

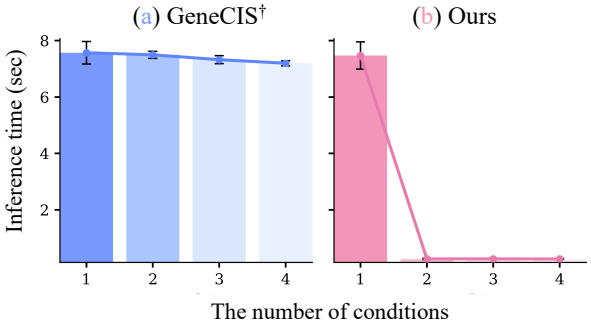


Figure 8. **Comparison of inference time on sampled CLAY-shoes dataset.** We compare the inference time between the symmetric similarity formulation of GeneCIS (*i.e.*, GeneCIS†), and ours. GeneCIS† shows high inference time whenever condition is given, while our method reduces after the first condition, benefiting from the decoupling feature extraction from the conditioning process.

Previous work [38] studied this rankability by identifying the rankable axis, CLAY can achieve this property naturally through conditional similarity space modulation. We also find that this rankable property holds in Clevr4-count, and provide in supplementary materials.

6. Conclusion

In this work, we introduce CLAY, a novel training-free conditional visual similarity computation method that adaptively modulates the fixed similarity of existing pre-trained VLMs to a text-conditional similarity. CLAY's adaptive conditioning strikes a sweet spot between accuracy and efficiency. These two factors are that prior methods struggled to achieve simultaneously, enabling practical and effective conditional retrieval. By leveraging the underlying hyperspherical geometry, our manifold-aware textual subspace enables theoretically grounded modeling of the conditional relationship between images. To support a comprehensive evaluation of the multi-faceted aspects of conditional retrieval, we construct a synthetic evaluation dataset containing diverse object and human images annotated with conceptual conditions. We believe this work opens up promising directions for building practical retrieval systems compliant to human intentions.

References

[1] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, 2023. 2, 3, 7

[2] 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. 4

[3] Black Forest Labs. Flux.1. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. 1-dev. 5

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

[5] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *ECCV*, 2020. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1

[7] 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

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[9] 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. 2, 3, 4

[10] Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *ICLR*, 2025. 4

[11] 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. 4

[12] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023. 1

[13] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *IJCV*, 124(2):237–254, 2017. 5

[14] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *CVPR*, 2024. 2, 3

[15] Søren Hauberg. Directional statistics with the spherical normal distribution. In *2018 21st international conference on information fusion (FUSION)*, 2018. 4

[16] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on graphics (TOG)*, 26(3):4–es, 2007. 2

[17] 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. 2, 3, 6

[18] Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. In *ECCV*, 2024. 2, 3

[19] 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. 5

[20] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *ICLR*, 2024. 2

[21] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *ICLR*, 2024. 3, 7

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

[23] Sehyun Kwon, Jaeseung Park, Minkyu Kim, Jaewoong Cho, Ernest K. Ryu, and Kangwook Lee. Image clustering conditioned on text criteria. In *ICLR*, 2024. 5

[24] Yangyang Li, Daqing Liu, Wu Liu, Allen He, Xinchen Liu, Yongdong Zhang, and Guoqing Jin. Omniprism: Learning disentangled visual concept for image generation. In *CoRR*, 2024. 5

[25] 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. 4

[26] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, 2021. 3

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

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

[29] Quang-Binh Nguyen, Minh Luu, Quang Nguyen, Anh Tran, and Khoi Nguyen. Csd-var: Content-style decomposition in visual autoregressive models. In *ICCV*, 2025. 2, 3, 4

[30] 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

[31] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, 2017. 2

[32] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 2

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

[34] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018. 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 6

[36] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *ICCV*, 2019. 5

[37] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *CVPR*, 2023. 2

[38] Ankit Sonthalia, Arnas Uselis, and Seong Joon Oh. On the rankability of visual embeddings. *arXiv preprint arXiv:2507.03683*, 2025. 8

[39] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *CVPR*, 2023. 2, 3, 5

[40] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *CVPR*, 2017. 2, 3

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

[42] Shijie Wang, Jianlong Chang, Haojie Li, Zhihui Wang, Wanli Ouyang, and Qi Tian. Open-set fine-grained retrieval via prompting vision-language evaluator. In *CVPR*, 2023. 3

[43] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE transactions on cybernetics*, 47(2):449–460, 2016. 2

[44] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. In *CVPR*, 2021. 3

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

[46] 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

[47] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 1, 2, 3, 6

[48] 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. 2, 3, 6, 7