Modeling Uncertainty in Composed Image Retrieval via Probabilistic Embeddings

Haomiao Tang[1*], **Jinpeng Wang**[1*†], **Yuang Peng**[1], **Guanghao Meng**[1], **Ruisheng Luo**[1], **Bin Chen**[2], **Long Chen**[3], **Yaowei Wang**[2,4], **Shu-Tao Xia**[1]
[1]Tsinghua University [2]Harbin Institute of Technology, Shenzhen
[3]The Hong Kong University of Science and Technology [4]Pengcheng Laboratory
thm23@mails.tsinghua.edu.cn ✉ wjp20@mails.tsinghua.edu.cn

Abstract

Composed Image Retrieval (CIR) enables users to search for images using multimodal queries that combine text and reference images. While metric learning methods have shown promise, they rely on deterministic point embeddings that fail to capture the inherent uncertainty in the input data, in which user intentions may be imprecisely specified or open to multiple interpretations. We address this challenge by reformulating CIR through our proposed **Co**mposed **P**robabilistic **E**mbedding (CoPE) framework, which represents both queries and targets as Gaussian distributions in latent space rather than fixed points. Through careful design of probabilistic distance metrics and hierarchical learning objectives, CoPE explicitly captures uncertainty at both instance and feature levels, enabling more flexible, nuanced, and robust matching that can handle polysemy and ambiguity in search intentions. Extensive experiments across multiple benchmarks demonstrate that CoPE effectively quantifies both quality and semantic uncertainties within Composed Image Retrieval, achieving state-of-the-art performance on recall rate. Code: https://github.com/tanghme0w/ACL25-CoPE.

1 Introduction

Composed Image Retrieval (CIR) (Vo et al., 2019) is a specialized task that enables image search by combining a reference image and a textual instruction that specifies desired modifications to the reference image. The instructions can range from attribute alterations (e.g., changing a shirt's logo) to contextual transformations (e.g., relocating a dog from indoors to outdoors). By leveraging visual and textual modalities in the query, CIR provides users with a powerful means to express complex search intentions, making it particularly valuable for applications in personalized content recommendation, e-commerce, and multimedia editing.

Recent approaches (Baldrati et al., 2022; Han et al., 2023; Wen et al., 2023; Yang et al., 2023; Xu et al., 2023; Baldrati et al., 2023; Saito et al., 2023) are mostly built upon metric learning, which aims to learn latent embeddings that effectively represent both multi-modal queries and candidate images. This approach is particularly promising as it enables models to leverage the expressive and pre-aligned embeddings from pre-trained models (Radford et al., 2021; Li et al., 2022, 2023a). However, existing CIR methods attempt to estimate accurate point embeddings in the latent space, which is intrinsically difficult due to two major challenges:

1) Imprecise or low-quality input. The model may encounter reference images that suffer from low resolution, blur, occlusion, and multiplicity. Meanwhile, modification texts may contain grammatical errors, colloquialisms, or perplexing sentences that does not convey sufficient information.

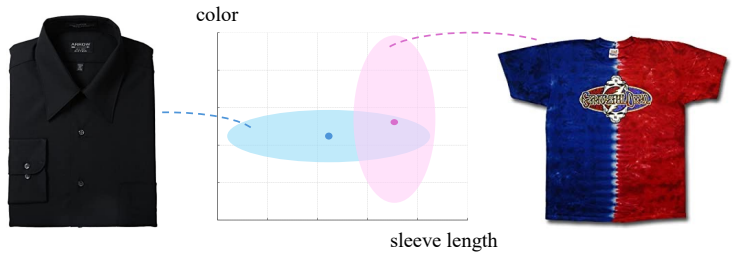
2) Ambiguous intentions. Text instructions like "make it more natural" or "improve the lighting" can have multiple valid interpretations. Image properties may also be subject to multiple perspectives. What one might consider "minimalist style" could be interpreted as "abstract style" by others, and a shirt that is considered long sleeve by some might be seen as medium sleeve for others.

The aforementioned data quality and semantic ambiguity issues manifest at varying degrees across different samples, introducing significant uncertainty to the model's input. Current point embedding approaches train models to minimize distances between positive pairs and maximize distances between negative pairs without accounting for the varying quality and ambiguity of individual samples. This uniform treatment of samples, regardless of their reliability, can lead to training instability and increased risk of over-fitting.

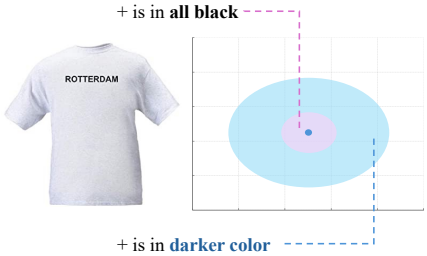
In this paper, we address these challenges by quantifying and reasoning about uncertainties for each individual input. To this end, we propose

*These authors contributed equally to this work.
†Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1210–1222



(a) **Visual ambiguity** the image on the left has more uncertainty in sleeve length and less uncertainty in color, while the image on the right has more uncertainty in color and less uncertainty in sleeve length.



(b) **Text ambiguity** "in darker color" is more ambiguous than "in all black", and thus should be encoded with larger uncertainty.

(c) **Image quality:** when searching for fashion images, the images on the right side are noisier in terms of background, angle, occlusion, irrelevant content, etc. And thus should be encoded with larger uncertainty.

Figure 1: Examples of uncertainty in composed image retrieval. Images are curated from Fashion-IQ dataset.

an uncertainty quantification framework that explicitly models query and candidate embeddings as probabilistic distributions rather than point estimates, with distribution variance reflecting the uncertainty of input data. Our probabilistic representation offers the following key advantages over deterministic point embeddings: 1) by capturing data quality through learning the overall magnitude of the variance, the model assigns different significance to different quality instances during the training and inference process. 2) through the learned distribution of variance across different feature dimensions for each instance, the model is aware of the semantic ambiguity in user's search intentions, reducing the importance of feature dimensions that are ambiguous and focus on features with clearer specification. Experiment results show that this approach improves training stability as well as overall recall performance.

We summarize our contributions as follows.

1. We propose CoPE, a novel probabilistic embedding approach to quantify the uncertainty and enhance CIR without additional annotations.

2. We develop an uncertainty-aware distance metric and systematic training objectives based on it, simultaneously modeling the magnitude and dimensional distribution of uncertainty.

3. We show CoPE effectively captures the uncertainty and achieves improved performance over deterministic methods on the CIR benchmark.

2 Related Work

2.1 Composed Image Retrieval

Research approaches in Composed Image Retrieval (CIR) generally fall into two main categories: **Supervised CIR** approaches (Vo et al., 2019; Kim et al., 2021; Dodds et al., 2020; Lee et al., 2021; Delmas et al., 2022; Couairon et al., 2022; Wang et al., 2022; Zhang et al., 2022; Zhao et al., 2022; Baldrati et al., 2022; Han et al., 2023; Wen et al., 2023; Yang et al., 2023; Xu et al., 2023) rely on triplet training data consisting of a reference image, modification text, and target image. These methods focus on developing sophisticated mechanisms to fuse the latent representations of reference image and text, as well as accurately capturing subtle visual differences between reference and target images. **Zero-shot CIR** methods (Baldrati et al., 2023; Saito et al., 2023; Tang et al., 2024; Lin et al., 2024; Suo et al., 2024) take a different approach by training on image-text paired datasets instead of triplets. They typically work by converting visual features from the reference image into pseudo-text tokens, effectively transforming the CIR task into a conventional text-to-image retrieval problem. While zero-shot approach enables training on more abundant data, the lack of explicit

triplet supervision often limits these methods' ability to interpret nuanced search intentions, resulting in notably lower performance compared to supervised approaches. In this paper, we mainly focus on the supervised setting of the CIR task.

Despite its significance, triplet data in supervised CIR is inherently noisy and ambiguous, as discussed in the previous section. Existing methods have attempted to mitigate this challenge through augmenting training data with additional high-quality samples (Jang et al., 2024; Gu et al., 2024; Feng et al., 2024; Ventura et al., 2024; Liu et al., 2024), developing fine-grained semantic parsing and decomposition schemes (Yang et al., 2024; Lin et al., 2024), leveraging large language models (LLMs) to refine or disambiguate user intentions (Baldrati et al., 2023; Karthik et al., 2024), or introduce regularization in the training process to improve the model's adaptability towards ambiguous inputs (Chen et al., 2024; Xu et al., 2024). Distinct from methods that attempts to *eliminate* the uncertainty in data, our proposed method *identifies* these uncertainties with probabilistic embedding, offering rich representation, stability in training and flexibility in matching.

2.2 Uncertainty Quantification

Uncertainty measures the degree of possibility that a model's prediction could be wrong. There are two fundamental factors of uncertainty within a model's prediction: (Kiureghian and Ditlevsen, 2009). i) *The imperfection of the model's capabilities*, and ii) *The deficiency of the input data*. The former, known as **Epistemic Uncertainty**, can be reduced by scaling up training data or improving model architecture. The latter, termed **Aleatoric Uncertainty**, arises from natural variations in data quality, ambiguity, and semantic content. Due to the infinite possible combinations of real-world data and its inherent randomness, aleatoric uncertainty cannot be eliminated even with additional training data (Kendall and Gal, 2017). In this paper we focus on the aleatoric uncertainty in CIR and provide insights for optimizing model performance under fixed data constraints. Specifically, we aim to develop a method to quantify the level of uncertainty for each data sample individually.

2.3 Probabilistic Embedding

Numerous works have explored the benefits of probabilistic embedding in tasks such as graph learning (Bojchevski and Günnemann, 2018), word embedding (Vilnis and McCallum, 2015), face recognition (Shi and Jain, 2019; Chang et al., 2020), 3D pose estimation (Sun et al., 2020), speaker diarization (Silnova et al., 2020), etc. Regarding tasks similar to ours, PFE (Shi and Jain, 2019) and DUL (Chang et al., 2020) discussed the application of probabilistic embedding in facial recognition. These methods leverage probabilistic embedding to identify the uncertainty in visual qualities of facial images. HIB (Oh et al., 2019) proposed a generalized soft contrastive framework to learn probabilistic embeddings. PCME (Chun et al., 2021) extends this framework to cross-modal retrieval. Its sequel (Chun, 2024) took a step further by introducing a closed form distance metric, improving the training stability and retrieval performance. Other works (Song and Soleymani, 2019; Andrei et al., 2022) have also applied probabilistic embedding to address the multi-matching or partial-matching problems. Different from the line of preceding works, our method not only learns **instance-wise** uncertainty, but also explicitly learns the **feature-wise** uncertainty. This is achieved by a unique distance metric that is aware of feature-wise uncertainty distribution and a hierarchical learning approach that enables multi-grained uncertainty modeling.

3 Method

3.1 Problem formulation

In the context of CIR, a model operates on triplet data $\{(x_r^i, x_t^i, x_c^i)\}_{i=1}^N$, where x_r^i denotes the i-th reference image, x_t^i denotes the i-th modification text, and x_c^i denotes their corresponding candidate image (i.e., target image). The model aims to produce multi-modal query representation $z_q^i = f_q(x_r^i, x_t^i)$ and candidate image representation $z_c^i = f_c(x_c^i)$ such that for a certain distance metric d, corresponding query-candidate pairs have smaller distances than non-corresponding pairs:

$$d(z_q^i, z_c^i) < d(z_q^i, z_c^j), \quad i \neq j \tag{1}$$

In our approach, we model each embedding as a distribution in the latent space characterized by a Gaussian prior: $z \sim \mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^D$ is the mean vector and $\Sigma \in \mathbb{R}^{D \times D}$ stands for the covariance matrix. For computational tractability, we assume mutual independence across embedding dimensions and constrain Σ to have non-zero entries only on its diagonal. Consequently, the above notation is reduced to $z \sim \mathcal{N}(\mu, \sigma^2 \mathrm{I})$,

where $\mu, \sigma^2 \in \mathbb{R}^D$ are both $D-$dimensional vectors and I is the $D \times D$ identity matrix.

The embedding model computes the distribution parameters for the reference image, the modification text, and the candidate image respectively:

$$\mu_t = f_T(x_t), \quad \sigma_t = g_T(x_t), \tag{2}$$

$$\mu_r = f_V(x_r), \quad \sigma_r = g_V(x_r), \tag{3}$$

$$\mu_c = f_V(x_c), \quad \sigma_c = g_V(x_c). \tag{4}$$

Here, $f_V(\cdot), g_V(\cdot)$ represent the mean and variance heads of the *vision* encoder branch, while $f_T(\cdot), g_T(\cdot)$ represent the mean and variance heads of the *text* encoder branch. The reference image and candidate image are processed through the same vision encoder (sharing $f_V(\cdot), g_V(\cdot)$) in order to leverage the visual priors of pre-trained backbones and reduce training complexity.

The query embedding is obtained by combining the Gaussian embeddings of the reference image and modification text through addition:

$$z_q = z_r + z_t \sim \mathcal{N}\left(\mu_r + \mu_t, \left(\sigma_r^2 + \sigma_t^2\right) \mathrm{I}\right). \tag{5}$$

This additive composition treats the text embedding as a displacement vector, representing the offset from the reference image embedding to the candidate image embedding within the latent space.

3.2 Model Architecture

3.2.1 Feature Extraction Network

We extract the mean components of probabilistic embeddings using the CLIP encoders (Radford et al., 2021). Considering that different text modification intentions emphasize different levels of granularity, we implement a multi-grained feature extraction mechanism that connects the lower layers of the CLIP vision transformer with the last hidden embedding layer. The weights of these connections are modulated the global text embedding through a Cross Attention (XA) gate

$$h_l' = \mathrm{LN}(h_l + \mathrm{XA}(h_l, \mu_t)), \tag{6}$$

$$h_N = \frac{1}{N} \sum_{l=1}^{N} h_l'. \tag{7}$$

$h_l \in \mathbb{R}^{L \times D}$ represents the hidden states at the l-th layer of the vision encoder. We notice that this modulator helps the model attend to different levels of representations for different instances, improving training stability as well as overall performance.

3.2.2 Uncertainty Quantification Head

The uncertainty quantification heads take the last hidden layer of the vision and text encoders as input and process them through parallel pathways to estimate both mean and variance parameters of the probabilistic embeddings. The network processes the last hidden embeddings through an MLP followed by local attention. The outputs are combined through a residual connection and fed into a Generalized Pooling Operator (GPO) to produce the final uncertainty estimates σ_r/σ_c. Uncertainties are captured independently for both modalities before being combined for the final retrieval task. Following (Chen et al., 2021), we implement the GPO as our final pooling operator, as it has been shown to effectively aggregate uncertainty information while maintaining probabilistic interpretability.

3.3 Uncertainty Learning

3.3.1 Instance-Wise Uncertainty Learning

At the instance level, uncertainty stems from the fact that data instances vary in qualities. In CoPE, this is represented by the magnitude of the uncertainty parameter $||\sigma||_2^2$, larger uncertainty values indicate less reliable instances.

We apply an batch-wise uncertainty-aware contrastive loss to enable the learning of overall instance representations, the form of which is is inspired by SigLIP (Zhai et al., 2023):

$$\mathcal{L}_{ij} = -\log \mathcal{S}\left(m_{ij}(-a \cdot d(z_i, z_j) + b)\right), \tag{8}$$

$$\mathcal{L}_C = \frac{1}{|B|} \sum_{i \in B} \sum_{j \in B} \mathcal{L}_{ij}. \tag{9}$$

Here, $\mathcal{S}$ is the Sigmoid function. m_{ij} is the label for a given query and candidate input, which equals 1 if they are matched and -1 otherwise. a, b are learnable parameters in which $a > 0$. $d(z_i, z_j)$ is the distance metric defining the matching behavior under the probabilistic embedding scheme. Given a pair of probabilistic embeddings $z_1 \sim \mathcal{N}\left(\mu_1, \sigma_1^2 \mathrm{I}\right)$ and $z_2 \sim \mathcal{N}\left(\mu_2, \sigma_2^2 \mathrm{I}\right)$, CoPE computes their distance via a closed-form distance metric

$$d(z_1, z_2) = \|\mu_1 - \mu_2\|_2^2 + \|\sigma_1 - \sigma_2\|_2^2 + 2D\bar{\sigma}_1\bar{\sigma}_2, \tag{10}$$

in which D is the embedding dimension, $\bar{\sigma}$ denotes the mean value across all dimensions of σ. The distance metric consists of two parts. The first part $\|\mu_1 - \mu_2\|_2^2 + \|\sigma_1 - \sigma_2\|_2^2$ is the Wasserstein-2 distance between the two Gaussian distributions,

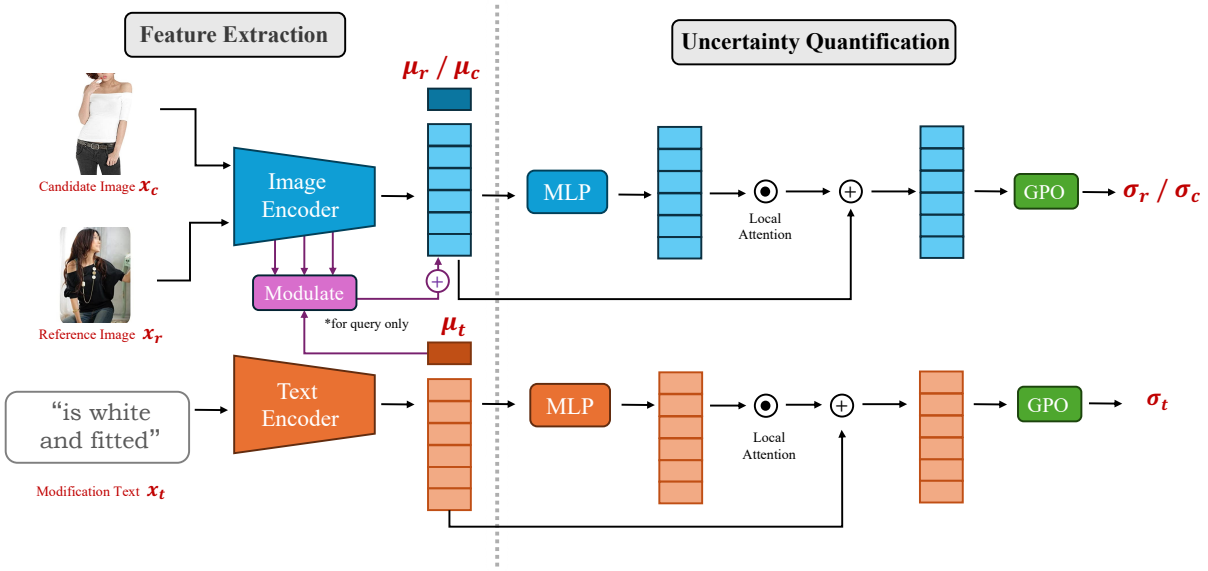


Figure 2: Overview of the Model Architecture. Embedding mean μ is derived by the global features of image and text encoders directly, while uncertainty values σ are produced by the uncertainty quantification heads, each consisting of a residual local attention branch and a generalized pooling operator (GPO). The visual branches for reference and candidate images share the same weights. A modulator is applied to enhance multi-grained feature extraction during query construction. All parameters are jointly optimized during the uncertainty learning process.

and the second part $D\bar{\sigma_1}\bar{\sigma_2}$ penalizes the matching between instances of high uncertainties. This distance metric bears some basic desirable properties, including i) *Non-negativity*. ii) *Symmetry*. iii) *Discernibility*, i.e., the closest embedding to an instance is always itself. To further illustrate the properties of this distance metric, we rewrite Equation (10) with respect to the Pearson Correlation Coefficient ρ between the uncertainties σ_1 and σ_2.

$$d(z_i, z_2) = \|\mu_1 - \mu_2\|_2^2 + \|\sigma_1\|_2^2 + \|\sigma_2\|_2^2 - 2\rho \cdot \text{std}(\sigma_1) \cdot \text{std}(\sigma_2). \quad (11)$$

This distance metric is positively proportional to the magnitude of both uncertainty terms $||\sigma_1||_2^2$ and $||\sigma_2||_2^2$, while being negatively proportional to the Pearson correlation coefficient ρ between the uncertainties. This formulation defines two important matching behaviors:

1) Higher overall uncertainty magnitudes yield larger distance. Given a query with comparable distances to multiple candidates, the model prioritizes the matches with lower uncertainty values, effectively favoring embeddings that represent more reliable and higher quality instances.

2) Similar uncertainty patterns across different feature dimensions yield smaller distance. When both embeddings share high uncertainty in some particular dimensions, those uncertainties are mutually canceled and contribute less to the overall distance. For instance, if a query that searches for a dress does not specify any information about sleeve length, then the uncertainty of sleeve length in the candidate images would also be ignored during the matching process. When uncertainties are uniform across all dimensions (i.e., low std(σ_1)) and std(σ_2)), the correlation term's impact diminishes, and the absolute magnitudes dominates.

The theoretical grounds of the loss function can be confirmed by a gradient analysis approach similar to that of (Chun et al., 2021). In essence, the proposed loss is in effect a weighting mechanism in which the model conducts uncertainty-aware matching by assigning higher weights to pairs with smaller distances for positive matches and to pairs with larger distances for negative matches. This creates a robust learning framework that does not severely penalize incorrect similarity predictions, but encodes the dissimilarity information in a more flexible variance span, thereby encouraging the model to learn the nuanced matching relationships while maintaining diversity in its embedding space.

3.3.2 Feature-Wise Uncertainty Learning

Feature-wise uncertainty learning operates on the basic assumption that different dimensions of the feature embedding represent different concepts, e.g., color, shape, texture. Feature-wise uncertainty

Table 1: Comparison with existing methods on Fashion-IQ dataset.

	Dress		Shirt		Top&tee		Avg		
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	Avg.
TIRG (Vo et al., 2019)	14.87	34.66	18.26	37.89	19.08	39.68	17.40	37.41	27.41
CIRPLANT (Liu et al., 2021)	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	30.20
CoSMo (Lee et al., 2021)	25.64	50.30	24.90	49.18	29.21	57.46	26.58	52.31	39.45
ARTEMIS (Delmas et al., 2022)	27.16	52.40	21.78	43.64	29.20	53.83	26.05	49.96	38.00
CompoDiff (Gu et al., 2024)	**40.65**	57.14	36.87	57.39	43.93	61.17	40.48	58.57	49.53
DWC (Yang et al., 2024)	32.67	57.96	35.53	60.11	40.13	66.09	36.11	61.39	48.75
CLIP4CIR (Baldrati et al., 2022)	33.81	59.40	39.99	60.45	41.41	65.37	38.40	61.74	50.07
SSN (Yang et al., 2024)	34.36	60.78	38.13	61.83	44.26	69.05	38.92	63.89	51.40
SADN (Wang et al., 2024)	40.01	65.10	43.67	66.05	48.04	70.93	43.91	67.36	55.63
CoPE (Ours)	39.85±0.30	**66.98**±0.34	**45.03**±0.39	**66.81**±0.31	**48.61**±0.49	**72.01**±0.34	**44.50**±0.40	**68.60**±0.36	**56.55**±0.40

Table 2: Comparison with existing methods on CIRR dataset.

	Recall@K				Recall$_{subset}$@K			$\frac{\text{R@5+R}_s\text{@1}}{2}$
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
CIRPLANT (Liu et al., 2021)	19.55	52.55	68.39	92.38	39.2	63.03	79.49	45.88
CompoDiff (Gu et al., 2024)	32.39	57.61	77.25	94.61	67.88	85.29	94.07	62.75
CLIP4CIR (Baldrati et al., 2022)	38.53	69.98	81.86	95.93	68.19	85.64	94.17	69.09
BLIP4CIR (Liu et al., 2024)	40.15	73.08	83.88	96.27	72.10	88.27	**95.93**	72.59
SSN (Yang et al., 2024)	43.91	77.25	86.48	97.45	71.76	88.63	95.54	74.51
SADN (Wang et al., 2024)	44.27	78.1	87.71	97.89	**72.71**	**89.33**	95.38	75.41
CoPE (Ours)	**49.18**±0.26	**80.65**±0.21	**89.86**±0.12	**98.05**±0.14	72.34±0.23	88.65±0.16	95.30±0.11	**76.49**±0.22

measures the different levels of ambiguity for different concepts within an instance. In CoPE, this is represented by the value distribution of σ across different feature dimensions.

The ambiguity of an instance can be estimated through a neighborhood analysis. Intuitively, if certain features of an embedding are ambiguous, their corresponding dimensions are expected to exhibit higher variance within the embedding's neighborhood. To capture this, we introduce the **neighborhood deviation loss** $\mathcal{L}_{ND}$, which enforces proportionality between the embedding uncertainty and the neighborhood feature deviation

$$\mathcal{L}_{ND} = \sum_{x \in \{r,t,c\}} \|\sigma_x - \frac{1}{K} \cdot \textbf{std}(\mu_{N_x})\|_2^2, \quad (12)$$

where $\{r, t, c\}$ represent the reference image, modification text, and candidate image, respectively. K is the number number of samples in the neighborhood. By minimizing this loss, the model enforces a proportional relationship between the embedding uncertainties (σ_x) and the standard deviation of neighborhood features **std**(μ_{N_x}). The k-NN search is conducted based on the mean feature embeddings μ_r, μ_t, μ_c using cosine similarity metric .

The overall loss is the combination of the aforementioned instance-wise contrastive loss and feature-wise neighborhood deviation loss, weighted by a parameter λ

$$\mathcal{L} = \mathcal{L}_C + \lambda \mathcal{L}_{ND}. \quad (13)$$

During implementation, we set λ to 0.2 for the best performance. A detailed study on this hyperparameter can be found in section 4.3.2

4 Experiment

4.1 Experiment Setting

Datasets and Metrics. We use two standard datasets for experiments: **Fashion-IQ** (Wu et al., 2021) is a fashion dataset containing 18,000 training triplets and 6,016 validation triplets, with 15,536 candidate images in total as candidate for validation. We report model performance on this dataset via Recall@K metric under K=10 and K=50, respectively. **CIRR** (Liu et al., 2021) comprises 36,554 image triplets derived from 21,552 real-world photographs originally sourced from NLVR2. While CIRR employs traditional Recall@K metrics similar to Fashion-IQ, it introduces an innovative evaluation framework called Recall$_{subset}$@K. This framework challenges models to identify target images within small groups of six visually similar images, thereby testing their capacity for fine-grained discrimination.

Implementation Details. CoPE employs CLIP-ViT-L/14 as the backbone. Training is conducted on a single A100-80G GPU with a batch size of

Table 3: Ablation study of CoPE on Fashion-IQ dataset.

	Dress		Shirt		Top&tee		Avg		
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	Avg.
H.C.	35.53	58.90	40.07	61.87	41.75	66.71	39.12	62.49	50.81
CoPE Ins.	37.42	63.08	41.90	61.06	42.46	65.27	40.59	63.14	51.87
CoPE Ins. + Feat. C	37.57	63.31	41.54	61.75	45.51	68.33	41.54	64.46	53.00
CoPE Ins. + Feat. R	39.85	65.94	43.20	63.48	44.36	66.43	42.47	65.28	53.88
CoPE Ins. + Feat. T	37.67	63.58	41.51	61.16	43.63	66.45	40.94	63.73	52.33
CoPE Ins. + Feat. R&C	38.44	**67.81**	44.33	63.61	44.33	67.38	42.37	66.27	54.32
CoPE Ins. + Feat. R&T	39.08	66.19	**45.11**	64.12	45.54	66.69	43.24	65.67	54.46
CoPE	**39.85**	66.98	45.03	**66.81**	**48.61**	**72.01**	**44.50**	**68.60**	**56.55**

128 and an initial learning rate of 2×10^{-6}. We implement an AdamW optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1.0 \times 10^{-7}$. We apply an Exponential Moving Average (EMA) strategy with an update rate of $r = 0.99$. To further enhance the learning process, we incorporate five distinct augmentation techniques to both reference and target images: cutout, HSV modification, rotation, scaling, and Gaussian noise addition. Each augmentation method is independently applied with a probability of 0.2 per image.

4.2 Comparisons with Existing Methods

We present a comparative evaluation of CoPE against existing CIR approaches, as shown in Tables 1 and 2. To robustly assess performance, we conduct 10 training runs with varying random seeds and report both the mean and standard deviation. Our results indicate that CoPE outperforms all baselines, including methods that utilize additional data sources (Gu et al., 2024). This provides strong empirical support for the advantage of our probabilistic embedding framework.

We notice a slight performance degradation on the Recall$_{subset}$@K setting of the CIRR dataset. This is likely due to the fact that the probabilistic embedding scheme works by hedging the matching process through uncertainty estimation, which may slightly obscure the difference between instances that are highly similar. Despite this drawback, our model still yields competitive performance and achieves overall SOTA on the CIRR dataset.

4.3 Model Analysis

4.3.1 Ablation Study

We illustrate the effectiveness of instance-wise and element-wise uncertainty learning via ablation study on Fashion-IQ dataset. As is shown in Table 3. *H.C.* Refers to training the model using conventional point embedding and hard contrastive loss (He et al., 2020). *CoPE-Ins* and *CoPE-Feat* means to apply the instance-wise or feature-wise uncertainty loss. *CoPE-Feat R, T, C* represents feature-wise uncertainty loss on the reference image, text, and candidate image respectively. All training settings will proceed until the training loss converges. Results show that **I.** model trained on the conventional hard contrastive loss performs far worse than models trained on probabilistic embedding. Moreover, we observe that the hard contrastive scheme quickly over-fits and stops improving after 10-12 epochs, while the performance of CoPE persistently optimizes even after more than 20 epochs, proving our advantage in training stability. **II.** instance-wise uncertainty loss and feature-uncertainty loss combined achieves better performance than using the two losses individually, thus proving the effectiveness of our training objectives.

4.3.2 Hyper-parameter Sensitivity

In this section we analyze the performance of CoPE under different hyper-parameter settings.

Effect of the neighborhood size. K determines the size of the neighborhood during feature-wise uncertainty learning. A larger K means considering more neighbors when computing feature-wise uncertainty, which can provide more comprehensive context but may also introduce noise from less relevant samples. As shown in Table 4, we experimented with different values of K ranging from 5 to 128. The results demonstrate that a moderate K (*e.g.*, 10) achieves better performance across metrics. When K becomes larger, performance begins to degrade, likely due to the inclusion of less relevant neighbors in the uncertainty estimation. Similarly, a too small neighborhood ($K = 5$) provides insufficient context, resulting in suboptimal performance. These findings suggest that setting K at approximately 10 strikes an effective balance be-

tween capturing enough neighborhood information and avoiding noise from distant samples.

Table 4: Model performance (average recall) on Fashion-IQ dataset under different settings of K.

	R@10	R@50	Overall
$K = 5$	43.56	66.18	54.87
$K = 10$	**44.50**	**68.60**	**56.55**
$K = 20$	43.94	67.85	55.90
$K = 50$	42.98	68.03	55.51
$K = 128$	42.39	67.77	55.08

Effect of the loss coefficient. As specified in Equation 13, λ defines the ratio between instance-wise uncertainty learning and element-wise uncertainty learning. A higher λ imposes a greater significance in the effect of feature-wise learning.

Table 5: Model performance (average recall) on Fashion-IQ dataset under different settings of λ.

	R@10	R@50	Overall
$\lambda = 1 \times 10^{-2}$	44.36	68.21	56.29
$\lambda = 2 \times 10^{-2}$	**44.50**	**68.60**	**56.55**
$\lambda = 5 \times 10^{-2}$	42.55	65.90	54.23
$\lambda = 1 \times 10^{-1}$	40.18	63.31	51.75

4.3.3 Understanding the Effectiveness of Uncertainty Modeling

Instance-level Uncertainty. We divide the validation samples from the Fashion-IQ dataset into 10 bins based on the overall magnitude of their query uncertainty, $||\sigma||_2^2$. For each subset, we compute the recall rate and plot the average recall rate against the uncertainty levels, as illustrated in Figure 3. The results reveal a clear decline in recall rate as uncertainty increases, demonstrating that the learned uncertainty effectively reflects the quality of individual data instances. Additionally, the negative correlation between recall rate and uncertainty is more pronounced in the R@50 setting compared to R@10, indicating that our uncertainty modeling method provides greater advantage in scenarios involving more ambiguous searches, i.e., with higher numbers of retrieval attempts.

Feature-level Uncertainty. Figure 4 shows a case study regarding the top instances with the highest or lowest uncertainty values on certain dimensions. CoPE model effectively identifies the semantic ambiguity in both texts and images.

4.3.4 Computational Efficiency

Unlike conventional cosine similarity or single-vector L2-distance search, our proposed distance

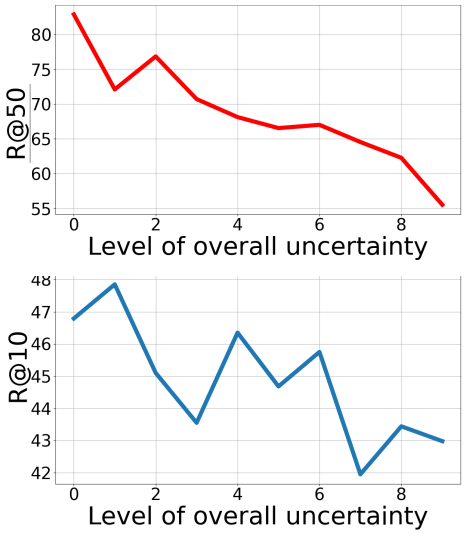


Figure 3: Fashion-IQ Average Recall with Respect to the Overall Level of Uncertainty.

metric (10) combines two L2 distance components with an additional inner product term. While this multi-term formulation offers uncertainty-aware distance modeling and enhances matching robustness, it also poses challenges for standard approximate nearest neighbor (ANN) search frameworks such as FAISS (Johnson et al., 2019), which lack native support for multi-vector fusion or hybrid similarity computations. Fortunately, advanced vector search systems such as Milvus (Wang et al., 2021) provide flexible indexing schemes and hybrid-field retrieval capabilities that align well with our formulation. To evaluate this, we implement our distance metric and a standard cosine similarity baseline using Milvus, and assess performance across databases of varying scales containing randomly generated dummy vectors.

Specifically, we decompose our metric into two components: a concatenated vector $[\mu; \sigma]$, stored in a `FLOAT_VECTOR` field and indexed via standard L2-based ANN structures (e.g., `IVFPQ`); and a scalar statistic $\bar{\sigma}$, stored separately and indexed for inner product computation (e.g., `IVF_FLAT`). Retrieval is conducted using Milvus' `hybrid_search` mechanism, with equally weighted sub-queries executed via `AnnSearchRequest` and `RRFRanker`.

As shown in Table 6, our CoPE embedding introduces some overhead compared to standard 768-dimensional embeddings due to its multi-vector nature. Nevertheless, when powered by hybrid vector search frameworks such as Milvus, the average retrieval time remains efficient even at large scale,

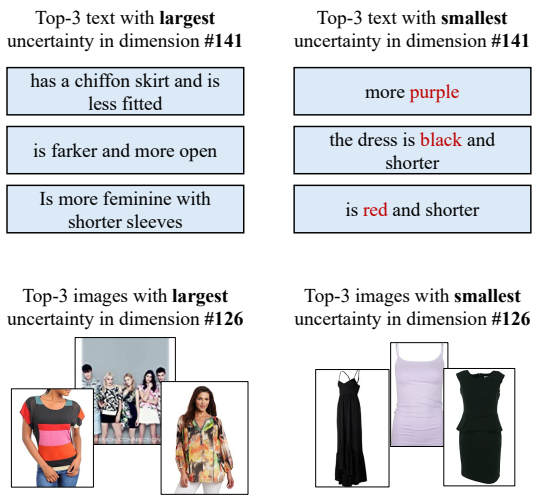


Figure 4: A case study on the uncertainty values across different dimensions of the embedding. Texts with high uncertainty in dimension 141 do not specify any color property, while texts with low uncertainty in dimension 141 all have a strong indication in color. The phenomenon holds for images in dimensions 126.

demonstrating the practicality of our method for real-world deployment.

Table 6: Average retrieval time per query (in milliseconds) across vector databases of varying size.

Method	**10K**	**100K**	**1M**	**10M**
Cosine Similarity	2.817	5.985	38.471	339.719
Cosine Similarity + Milvus	**0.202**	**2.076**	**2.496**	**5.498**
CoPE	5.474	12.535	115.945	734.305
CoPE + Milvus	3.230	5.589	13.601	46.733

4.3.5 Compatibility with Other Backbones

Our uncertainty learning process is independent of the feature extraction backbone and its pre-training process. To prove this, we provide supplementary experiments with the SigLIP backbone on the Fashion-IQ dataset.

Table 7: Retrieval performance on Fashion-IQ under different visual backbones.

Method	Mean R@10	Mean R@50	Mean All
Hard Contrastive (CLIP)	39.12	62.49	50.81
Hard Contrastive (SigLIP)	40.88	63.29	52.09
CoPE (CLIP)	44.50	68.60	56.55
CoPE (SigLIP)	46.51	70.02	58.27

Results show that the performance gain of CoPE does not rely on a specific backbone, but stems from its inherent modeling of feature uncertainty.

5 Conclusion

We address the challenge of data uncertainty in compositional image retrieval (CIR) through our proposed **Co**mposed **P**robabilistic **E**mbedding (CoPE) framework, which represents both queries and targets as Gaussian distributions in latent space. Through carefully designed distance metrics and hierarchical learning objectives, CoPE explicitly captures uncertainty at both instance and feature levels, enabling more nuanced and robust matching behavior under diverse compositional queries.

Broader Impact. We believe our work has broader implications across both the NLP and multimodal learning communities (Han et al., 2024; Zhao et al., 2023a; Wei et al., 2024; Yu et al., 2025; Wei et al., 2025). Our uncertainty-aware perspective aligns closely with ongoing research into semantic ambiguity and underspecified interpretations. The probabilistic modeling of meaning can be helpful to tasks such as word sense disambiguation (Bevilacqua et al., 2021), semantic role labeling (Li et al., 2023b), and cross-lingual representation learning (Gao et al., 2023).

Our method also holds promise for other advancing personalized retrieval and generation tasks (Gal et al., 2022; Ruiz et al., 2022; Wang et al., 2023; Zhao et al., 2023b; Peng et al., 2025; Lu et al., 2025; Meng et al., 2025; Dong et al., 2023; Liu et al., 2025; Xie et al., 2024; Tan et al., 2024; Zhou et al., 2025), where modeling uncertainty is critical for interpreting and aligning with highly subjective or ambiguous user intent. We believe the probabilistic approach provides a principled and extensible foundation for such applications, and we anticipate its adaptation to domains beyond CIR.

A Limitations

While the CoPE framework offers notable advances in compositional image retrieval by modeling uncertainty in both queries and targets, several technical limitations remain. First, we observe that although our model consistently achieves higher overall performance, it tends to exhibit stronger gains on broader retrieval metrics (e.g., Recall@50) than on stricter top-ranked evaluations (e.g., Recall@10 or Recall on small subsets). This suggests a potential trade-off between recall coverage and ranking sharpness, which may be addressed through more fine-grained contrastive alignment or dynamic re-ranking mechanisms. Second, com-

pared to conventional cosine similarity-based methods, our uncertainty-aware distance function introduces minor computational overhead due to its multi-term formulation and multi-vector indexing. Lastly, our current training strategy assumes a static uncertainty profile per image-query pair. Future work may explore context-dependent or task-aware uncertainty adaptation, enabling better personalization and generalization across retrieval domains.

Acknowledgments

We thank the anonymous reviewers and chairs for their efforts and constructive suggestions. This work is supported in part by the National Natural Science Foundation of China under grants 624B2088, 62171248, and 62301189; the PCNL KEY project (PCL2023AS6-1); Shenzhen Science and Technology Program under Grant KJZD20240903103702004, JCYJ20220818101012025, and GXWD20220811172936001. Long Chen was supported by the Hong Kong SAR RGC Early Career Scheme (26208924), the National Natural Science Foundation of China Young Scholar Fund (62402408), Huawei Gift Fund, and the HKUST Sports Science and Technology Research Grant (SSTRG24EG04).

References

Neculai Andrei, Yanbei Chen, and Zeynep Akata. 2022. Probabilistic Compositional Embeddings for Multimodal Image Retrieval. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4546–4556.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-Shot Composed Image Retrieval with Textual Inversion. In *IEEE International Conference on Computer Vision (ICCV)*, pages 15292–15301.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pages 4330–4338.

Aleksandar Bojchevski and Stephan Günnemann. 2018. Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking. In *International Conference on Learning Representations (ICLR)*.

Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. 2020. Data Uncertainty Learning in Face Recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5709–5718.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the Best Pooling Strategy for Visual Semantic Embedding. In *Computer Vision and Pattern Recognition (CVPR)*, pages 15789–15798.

Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. 2024. Composed Image Retrieval with Text Feedback via Multi-grained Uncertainty Regularization. *The Twelfth International Conference on Learning Representations*.

Sanghyuk Chun. 2024. Improved Probabilistic Image-Text Representations. In *The Twelfth International Conference on Learning Representations*.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic Embeddings for Cross-Modal Retrieval. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8415–8424.

Guillaume Couairon, Matthijs Douze, Matthieu Cord, and Holger Schwenk. 2022. Embedding Arithmetic of Multimodal Queries for Image Retrieval. In *CVPR Workshops*.

Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. 2022. Artemis: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity. In *International Conference on Learning Representations (ICLR)*.

Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. 2020. Modality-Agnostic Attention Fusion for visual search with text feedback. *arXiv*, abs/2007.00145.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*.

Zhangchi Feng, Richong Zhang, and Zhijie Nie. 2024. Improving Composed Image Retrieval via Contrastive Learning with Scaling Positives and Negatives. In *ACM Multimedia 2024*, volume abs/2404.11317.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. Learning multilingual sentence representations with cross-lingual consistency regularization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 269–278.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. 2024. Compodiff: Versatile Composed Image Retrieval With Latent Diffusion. *Transactions on Machine Learning Research (TMLR)*, 2024.

Chunrui Han, Jinrong Yang, Jianjian Sun, Zheng Ge, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. 2024. Exploring recurrent long-term temporal fusion for multi-view 3d perception. *IEEE Robotics and Automation Letters*.

Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. 2023. Fame-ViL: Multi-Tasking Vision-Language Model for Heterogeneous Fashion Tasks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2669–2680.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735. IEEE.

Young Kyun Jang, Donghyun Kim, Zihang Meng, Dat Huynh, and Ser-Nam Lim. 2024. Visual Delta Generator with Large Multi-modal Models for Semi-supervised Composed Image Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. 2024. Vision-by-Language for Training-Free Compositional Image Retrieval. In *The Twelfth International Conference on Learning Representations*.

Alex Kendall and Yarin Gal. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 5574–5584.

Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. 2021. Dual Compositional Learning in Interactive Image Retrieval. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1771–1779.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112.

Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. Cosmo: Content-Style Modulation for Image Retrieval With Text Feedback. In *Computer Vision and Pattern Recognition (CVPR)*, pages 802–812.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. Blip-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning (ICML)*, pages 19730–19742.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. Blip: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900.

Tao Li, Ghazaleh Kazeminejad, Susan Windisch Brown, Vivek Srikumar, and Martha Palmer. 2023b. Learning semantic role labeling from compatible label sequences. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1489–1503.

Haoqiang Lin, Haokun Wen, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. 2024. Fine-grained Textual Inversion Network for Zero-Shot Composed Image Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 34, pages 240–250. ACM.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. 2025. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*.

Zheyuan Liu, Cristian Rodriguez Opazo, Damien Teney, and Stephen Gould. 2021. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2105–2114.

Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. 2024. Bi-directional Training for Composed Image Retrieval via Text Prompt Learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5741–5750.

Xingyu Lu, Jinpeng Wang, Jieming Zhu, Zhicheng Zhang, Deqing Zou, Hai-Tao Zheng, Shu-Tao Xia, and Rui Zhang. 2025. Roma: Recommendation-oriented language model adaptation using multi-modal multi-domain item sequences. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Guanghao Meng, Sunan He, Jinpeng Wang, Tao Dai, Letian Zhang, Jieming Zhu, Qing Li, Gang Wang, Rui Zhang, and Yong Jiang. 2025. Evdclip: Improving vision-language retrieval with entity visual descriptions from large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6126–6134.

Seong Joon Oh, Kevin P. Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew C. Gallagher. 2019. Modeling Uncertainty with Hedged Instance Embeddings. In *International Conference on Learning Representations (ICLR)*.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge,

Xiangyu Zhang, and Shu-Tao Xia. 2025. Dreambench++: A Human-Aligned Benchmark for Personalized Image Generation. In *The Thirteenth International Conference on Learning Representations*, volume abs/2406.16855.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and I. Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, pages 8748–8763.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval. In *Computer Vision and Pattern Recognition (CVPR)*, pages 19305–19314.

Yichun Shi and Anil K. Jain. 2019. Probabilistic Face Embeddings. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6901–6910.

Anna Silnova, Niko Brummer, Johan Rohdin, Themos Stafylakis, and Lukás Burget. 2020. Probabilistic Embeddings for Speaker Diarization. In *Speaker and Language Recognition Workshop (Odyssey)*, pages 24–31.

Yale Song and Mohammad Soleymani. 2019. Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1979–1988.

Jennifer J. Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. 2020. View-Invariant Probabilistic Embedding for Human Pose. In *European Conference on Computer Vision (ECCV)*, pages 53–70.

Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. 2024. Knowledge-Enhanced Dual-stream Zero-shot Composed Image Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuqi Tan, Yuang Peng, Hao Fang, Bin Chen, and Shu-Tao Xia. 2024. Waterdiff: Perceptual image watermarks via diffusion model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3250–3254. IEEE.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. 2024. Context-I2W: Mapping Images to Context-Dependent Words for Accurate Zero-Shot Composed Image Retrieval. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 5180–5188.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2024. Covr: Learning Composed Video Retrieval from Web Video Captions. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 5270–5279.

Luke Vilnis and Andrew McCallum. 2015. Word Representations via Gaussian Embedding. In *International Conference on Learning Representations (ICLR)*.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6439–6448.

Chao Wang, Ehsan Nezhadarya, Tanmana Sadhu, and Shengdong Zhang. 2022. Exploring Compositional Image Retrieval with Hybrid Compositional Learning and Heuristic Negative Mining. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1273–1285.

Jinpeng Wang, Ziyun Zeng, Yunxiao Wang, Yuting Wang, Xingyu Lu, Tianxiang Li, Jun Yuan, Rui Zhang, Hai-Tao Zheng, and Shu-Tao Xia. 2023. Missrec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6548–6557.

Junyu Wang, Yuchen Xia, Jian Yu, Xinyu Wang, Zhen Wu, Zhujie Li, Jiaxuan Yang, Yan Tang, Dong Jin, Xia Liu, et al. 2021. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD)*, pages 2693–2704. ACM.

Yifan Wang, Wuliang Huang, Lei Li, and Chun Yuan. 2024. Semantic Distillation from Neighborhood for Composed Image Retrieval. In *ACM Multimedia 2024*.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model.

Yana Wei, Liang Zhao, Kangheng Lin, En Yu, Yuang Peng, Runpei Dong, Jianjian Sun, Haoran Wei, Zheng Ge, Xiangyu Zhang, et al. 2025. Perception in reflection. *arXiv preprint arXiv:2504.07165*.

Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. 2023. Target-Guided Composed Image Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*. ACM.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogério Feris. 2021. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11307–11317.

Yuqiu Xie, Bolin Jiang, Jiawei Li, Naiqi Li, Bin Chen, Tao Dai, Yuang Peng, and Shu-Tao Xia. 2024. Gladcoder: stylized qr code generation with grayscale-aware denoising process. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7780–7787.

Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. 2023. Multi-Modal Transformer With Global-Local Alignment for Composed Query Image Retrieval. *IEEE Transactions on Multimedia*, 25:8346–8357.

Yahui Xu, Jiwei Wei, Yi Bin, Yang Yang, Zeyu Ma, and Heng Tao Shen. 2024. Set of Diverse Queries With Uncertainty Regularization for Composed Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):10494–10506.

Qu Yang, Mang Ye, Zhaohui Cai, Kehua Su, and Bo Du. 2023. Composed Image Retrieval via Cross Relation Network With Hierarchical Aggregation Transformer. *IEEE Transactions on Image Processing*, 32:4543–4554.

Xingyu Yang, Daqing Liu, Heng Zhang, Yong Luo, Chaoyue Wang, and Jing Zhang. 2024. Decomposing Semantic Shifts for Composed Image Retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):6576–6584.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. 2025. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. In *IEEE International Conference on Computer Vision (ICCV)*, pages 11941–11952.

Feifei Zhang, Ming Yan, Ji Zhang, and Changsheng Xu. 2022. Comprehensive Relationship Reasoning for Composed Query Based Image Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM.

Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, et al. 2023a. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *arXiv preprint arXiv:2307.09474*.

Minyi Zhao, Jinpeng Wang, Dongliang Liao, Yiru Wang, Huanzhong Duan, and Shuigeng Zhou. 2023b. Keyword-based diverse image retrieval by semantics-aware contrastive learning and transformer. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1262–1272.

Yida Zhao, Yuqing Song, and Qin Jin. 2022. Progressive Learning for Image Retrieval with Hybrid-Modality Queries. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1012–1021.

Deyu Zhou, Quan Sun, Yuang Peng, Kun Yan, Runpei Dong, Duomin Wang, Zheng Ge, Nan Duan, Xiangyu Zhang, Lionel M Ni, et al. 2025. Taming teacher forcing for masked autoregressive video generation. *arXiv preprint arXiv:2501.12389*.