Towards Robust Uncertainty Calibration for Composed Image Retrieval

Yifan Wang[1], **Wuliang Huang**[2], **Yufan Wen**[1], **Shunning Liu**[1], **Chun Yuan**[1]*
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Institute of Computing Technology, Chinese Academy of Sciences
{yifan-wa22@mails, wenyf24@mails, lsn24@mails, yuanc@sz}.tsinghua.edu.cn
huangwuliang19b@ict.ac.cn

Abstract

The interactive task of composed image retrieval aims to retrieve the most relevant images with the bi-modal query, consisting of a reference image and a modification sentence. Despite significant efforts to bridge the heterogeneous gap within the bi-modal query and leverage contrastive learning to reduce the disparity between positive and negative triplets, prior methods often fail to ensure reliable matching due to aleatoric and epistemic uncertainty. Specifically, the aleatoric uncertainty stems from underlying semantic correlations within candidate instances and annotation noise, and the epistemic uncertainty is usually caused by overconfidence in dominant semantic categories. In this paper, we propose *Robust UNcertainty Calibration* (**RUNC**) to quantify the uncertainty and calibrate the imbalanced semantic distribution. To mitigate semantic ambiguity in similarity distribution between fusion queries and targets, RUNC maximizes the matching evidence by utilizing a high-order conjugate prior distribution to fit the semantic covariances in candidate samples. With the estimated uncertainty coefficient of each candidate, the target distribution is calibrated to encourage balanced semantic alignment. Additionally, we minimize the ambiguity in the fusion evidence when forming the unified query by incorporating orthogonal constraints on explicit textual embeddings and implicit queries, to reduce the representation redundancy. Extensive experiments and ablation analysis on benchmark datasets FashionIQ and CIRR verify the robustness of RUNC in predicting reliable retrieval results from a large image gallery.

1 Introduction

The task of composed image retrieval (CIR) [1, 2, 3, 4, 5] is emerging as a multi-modal interaction form to accommodate the flexible search requirements. Distinguished from traditional single-modal image retrieval or cross-modal retrieval, the queries of CIR support two modalities, *i.e.*, reference images and modification texts that illustrate alternations on the reference images. To identify the most correlated images among the massive candidate images, the core objective is to establish semantic connections between target images and bi-modal queries through similarity measurement and bridge the heterogeneous modality gap across different modalities within queries. Exploring multi-modal data integration and natural interaction requirements on this task could provide underlying support for tasks such as visual question answering (VQA) [6, 7, 8], visual reasoning [9, 10, 11], etc. as the basis of multi-modal understanding [12].

One widely recognized challenge for composed image retrieval is the comprehension of the semantic conflict in the bi-modal query. Despite the modality gap in the query composition, the modifications

*Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

expressed by the query text have led to disagreements with the reference images, hindering the understanding of the multi-modal input and formation of a unified query representation. Existing

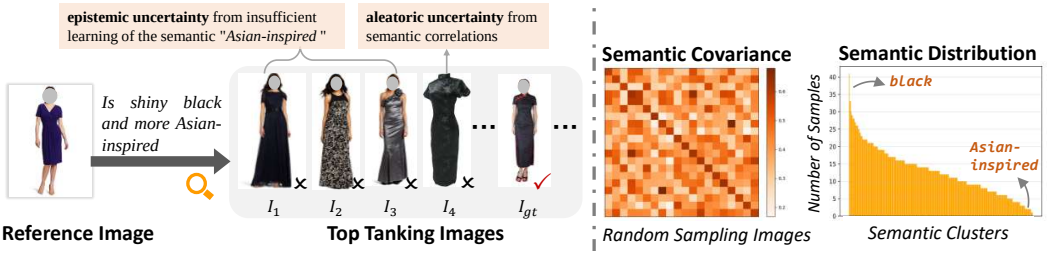


Figure 1: Illustration of uncertain matching results. (*left*) Unreliable top-ranking results disrupted by insufficient learning of "*Asign-inspired* " and partial semantic correlations. (*right*) Strong correlations within the semantic matrix and imbalanced semantic distributions underlying visual candidates.

works were devoted to mining the effective components within the queries by multi-granular feature matching, *e.g.*, word-level [13, 14], patch-level [15], token-level [16, 17], and hierarchical-level [18]. Inspired by the significant achievement of vision-language models on the massive corpus, recent models adopted transformer-based encoders in CLIP [19, 20] and BLIP [21, 22] to enhance visual and textual features [23] and align semantics in the query, owing to multiple inter- and intra-attention mechanisms to extract salient information. Besides, to capture the correspondences between multi-modal queries and targets, current approaches [24, 5] mainly project the query and target features in the joint space and measure pairwise similarities in the contrastive learning framework. This pipeline regards the CIR task as a classification task to distinguish between matched and mismatched triplets.

Contrastive learning essentially follows the assumption that there is no overlap between the distributions of positive and negative triplets. However, images inherently contain rich semantic concepts that cannot be explicitly assigned as independent labels, *i.e.*, negative candidates exhibit certain underlying semantic dependencies with ground-truth positive images. Fig. 1 illustrates that the annotated "negative" images I_4 and ground-truth images I_{gt} exhibit strong visual relations, while the semantic correlations remain disregarded within contrastive learning paradigm. Moreover, it is inevitable that not all images that satisfy the query requirements are labeled as positives, and the modification descriptions lack clarity in accurately conveying the intended image. The collaborative impacts of the above factors lead to *aleatoric uncertainty*, which is caused by the implicit semantics dependencies and noise labelling issues in the datasets. Additionally, semantic concept imbalance is also significant in images where specialized designs like "Asian-inspired" are limited in the fashion domain, as seen in Fig. 1 (*right*). Furthermore, statistical properties of widely-used softmax function in either attention mechanisms [25, 13, 23, 21] or cross-entropy calculation, further increase the biased estimation, leading to overconfident predictions dominated by the explicit semantic concepts and overlook of discriminative details (as shown in I_1, I_2, and I_3 in Fig. 1). The imbalanced distributions of semantic concepts in visual features and over-concentration on salient features result in *epistemic uncertainty*, which tends to overfit the majority semantics and result in low confidence when facing minority semantic categories.

To address the aforementioned challenges, we propose *Robust UNcertainty Calibration* (RUNC) to ensure credible semantic bridging between bi-modal queries and visual targets, as shown in Fig. 2. In order to perceive the latent semantic correlations and quantify uncertainty, we incorporate Normal Inverse Gamma distribution as evidential priors to fit the semantic covariances in the candidate images. Through maximizing the evidence by the model and imposing a penalty on the incorrect evidence, the inferred probabilistic distributions estimate aleatoric and epistemic uncertainty on each candidate image. With more emphasis on uncertain images with rare and ambiguous semantics, we assign uncertainty coefficients to calibrate the target distribution when supervising the query distribution. To further decrease the ambiguity when composing fusion representations for the coupled query images and texts and ensure the semantic consistency, implicit query embedding is introduced to drive the query embeddings to distill more retained visual semantics rather than redundant representations in text modifiers through aligning with the targets and orthogonal to query texts. Experimental results on widely-adopted datasets FashionIQ and CIRR verify that RUNC yields robust and reliable rankings.

In summary, the proposed RUNC makes the following contribution:

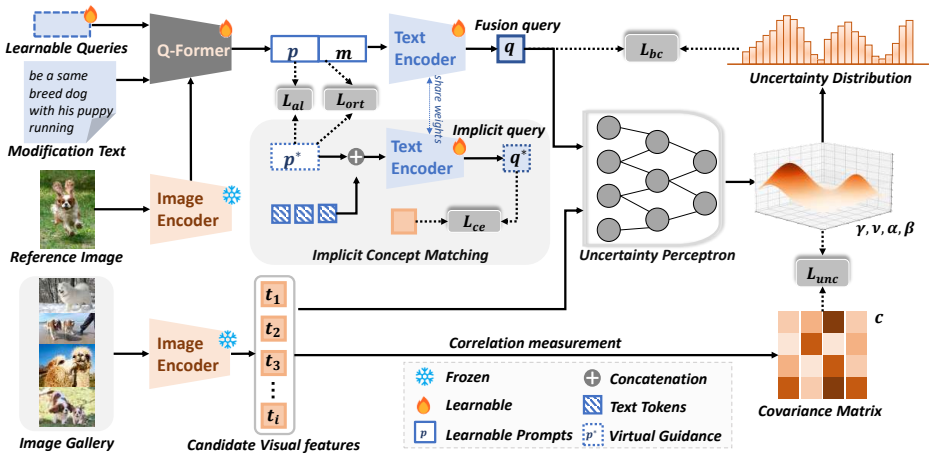


Figure 2: **The Framework of RUNC**. Uncertainty perceptron introduces evidential priors to fit the semantic covariance and yield uncertainty distribution to calibrate the supervision on the fusion query. Implicit guidances p^* are incorporated to distill effective features for retention and modification.

- We propose a novel uncertainty calibration approach to address the misguidance of dominant semantics and semantic overlaps when matching bi-modal queries with target images.
- We employ high-order evidential priors to quantify the aleatoric and epistemic uncertainty and adaptively adjust the semantic distribution imbalance based on uncertainty estimation.
- To minimize the ambiguity when fusing hybrid-modal queries, we assign orthogonal independent constraints on explicit textual embeddings and implicit queries to distill effective features for retention and modification.

2 Related Works

2.1 Composed Image Retrieval

To address the growing need for flexible retrieving with multi-modal data, composed image retrieval focuses on exploring the integration of dual-modal input of reference images and modification text and matching with the desired images. Existing works are dedicated to mitigating the heterogeneous gap of multi-modal data in the query domain by extracting effective semantic components across different modalities [1, 26, 27] via multiple projection layers [1], cross-modal attention [18, 28], and graph-based propagation [13, 29]. For instance, FashionVLP [30] introduced multi-layer self-attention on the combined tokens of visual regions from the reference image and words from modifiers. To facilitate correlations with query images and query texts, Bai *et al.* [31] incorporated sentence-level prompts with visual features and textual tokens in the Q-Former structure [32]. Through contrastive learning, the fusion queries are guided toward the target features and pushed away from other candidate features by comparing positive and negative samples. DCNet [33], ComposeAE [27], and CaLa [22] applied bi-directional constraints to strengthen the semantic consistency across references, modifiers, and targets. Recent advances [34, 16] reconstructed triplet data with higher-quality to ensure that contrastive learning effectively captures semantic alignment, which requires expensive annotation efforts. With a focus on the separation of positive and negative matching in the above approaches, the latent semantic correlations among the images, particularly false negatives, may hinder the actual semantic alignment. Distinguished from previous methods, this work deploys prior distributions to model the intrinsic covariance of candidate features and addresses the semantic imbalance by uncertainty calibration to enhance the robustness.

2.2 Uncertainty Estimation

Though deep learning models are currently excelling across various domains, most of them typically provide predictions without considering the confidence of the outcomes [35, 36]. The prediction uncertainty stemming from data noise, model overconfidence, and biased learning can significantly

impact the robustness of noisy labels, generalization on unseen classes, and model interpretability. Based on Bayesian Neural Networks (BNN) [37] and Monte Carlo Dropout, uncertainty models [38] estimated prediction variance through multiple forward propagation samples in out-of-distribution (OOD) scenarios [39, 40]. MGUR [4] applied weighted Gaussian noises on whitened features to simulate data jitter and one-to-many correspondences. Subjective Logic (SL) formalized the concept of belief assignment in the Dempster-Shafer theory of evidence as a Dirichlet distribution [41, 42] to quantify the belief mass and uncertainty. Deep Evidential Regression (DER) [43, 44] introduced high-order priors to capture the confidences supporting the model prediction. Evidential deep learning avoided the computational bottlenecks of traditional Bayesian methods by obtaining uncertainty in a single forward inference. In this work, the proposed RUNC extends the uncertainty estimation to explore latent correlations in the coupling visual semantics and calibrate the imbalanced semantic distributions when aligning composed queries and targets.

3 Methodology

3.1 Problem Setting

The multi-modal dataset for composed image retrieval includes N query-to-target pairs. Each matched data $(\mathcal{I}_r, M, \mathcal{I}_t)$ contains one reference image $\mathcal{I}_r$, one modification sentence $\mathcal{M}$, and one target image $\mathcal{I}_t$. To bridge the modality gap between the visual and textual inputs, the retrieval pipeline utilizes pretrained encoders [19, 21] to project all the raw image inputs into the semantic latent space. We denote candidate visual features as $\{\boldsymbol{t}_i\}_{i=1}^{N}$ for N images in the gallery, where $\boldsymbol{t}_i \in \mathbb{R}^d$ and d denotes dimensions. In the branch of combining the cross-modal query, the lightweight Q-Former [32] is employed to obtain the interactive prompt embedding $\boldsymbol{p}$ along with text features $\boldsymbol{m}$, with the input of visual features from the original reference images, instructions from the modification text, and learnable queries. Afterward, the prompt embedding $\boldsymbol{p}$ is further projected in the same latent space as $\boldsymbol{t}$ to yield the fusion query representation $\boldsymbol{q}$.

The essence of the composed image retrieval task lies in accurately measuring the semantic distances between the queries and targets in the embedding space. With the aim of pushing the fusion query representations towards the target features $(\boldsymbol{q}_i \rightarrow \boldsymbol{t}_i)$ while separating query representations from irrelevant candidate samples $(\boldsymbol{q}_i \longleftrightarrow \boldsymbol{t}_j)$, existing frameworks [45, 21] mainly adopt the contrastive loss to classify positive pairs $(\boldsymbol{q}_i, \boldsymbol{t}_i)$ and negative pairs $(\boldsymbol{q}_i, \boldsymbol{t}_j | j \neq i)$:

$$\mathcal{L}_{cl} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\boldsymbol{s}_{ii}/\tau)}{\sum_{j=1}^{B}\exp(\boldsymbol{s}_{ij}/\tau)}, \tag{1}$$

where B is the batch size and τ refers to the temperature parameter. The matching score $\boldsymbol{s}_{ij}$ between composed query i and candidate feature j is computed based on the cosine distance $\boldsymbol{s}_{ij} = \frac{\boldsymbol{q}_i \cdot \boldsymbol{t}_j}{\|\boldsymbol{q}_i\| \|\boldsymbol{t}_j\|}$, where $\|\cdot\|$ means the L2 normalization.

Though cross-entropy loss is efficient in large-scale retrieval applications, it cannot be fully adapted to this complicated interactive retrieval task. In this context, contrastive loss is primarily focused on aligning the matching scores matrix $\boldsymbol{s} \in \mathbb{R}^{B\times B}$ with the diagonal matrix. Specifically, for normalized features, the matching value for the fusion query to the target features should ideally approach 1, while the matching value for all the other samples should be as close to 0 as possible. However, due to the strong semantic correlations among most candidate images in the image gallery and noise labels during the dataset construction process, non-diagonal negative samples may exhibit semantic overlap. Furthermore, the fused features derived from the complex multi-modal interactive inputs inherently introduce semantic uncertainty and instability, which results in loose and unstable semantic construction with traditional supervision loss. To this end, the uncertainty of alignments between the cross-modal query and candidate images is imperative to estimate to construct a robust retrieval model. As shown in Fig. 2, we deploy distribution-based uncertainty estimation to quantify the semantic ambiguity and bring aleatoric and epistemic uncertainty into consideration when setting the training objective.

3.2 Uncertainty Estimation

Priors of Semantic Covariance Matrix. As aforementioned, the high frequency of coupling semantic concepts (as shown in Figure 1 (*left*)) in the candidate images poses challenges to the

model's prediction of similarities between queries and targets. Using hard-coded supervision with binary values of 1 or 0 for positive and negative sample pairs respectively fails to accurately represent the real retrieval scenario. In the proposed RUNC, we approach the prediction of matching values in the retrieval process as a regression problem, assuming that the distribution of the semantic correlations conforms to Gaussian distributions and is independent and identically distributed (i.i.d). The correlations across various target semantic concepts are acquired by the interactions of candidate features in the target image space, to measure the potential semantic overlaps between different pairs.

$$\boldsymbol{c}_{ij} = \frac{\boldsymbol{t}_i \cdot \boldsymbol{t}_j}{\|\boldsymbol{t}_i\|\|\boldsymbol{t}_j\|}, \tag{2}$$

where $\boldsymbol{c}_{ij}$ stands for the semantic covariance score between the i-th sample and the j-th sample.

The objective of uncertainty estimation is to estimate a prior distribution to reconstruct the variance and mean of the Gaussian distribution of semantic concepts [46, 47, 43]. Therefore, a high-order Normal Inverse Gamma (NIG) prior is introduced to model the Gaussian output:

$$(\boldsymbol{c}_{1i}, \boldsymbol{c}_{2i}, ..., \boldsymbol{c}_{Ni}) \sim \mathcal{N}(\mu_i, \delta_i^2), \quad \mu_i \sim \mathcal{N}(\gamma_i, \delta_i^2 \nu_i^{-1}), \quad \delta_i^2 \sim \Gamma^{-1}(\alpha_i, \beta_i), \tag{3}$$

where $\Gamma(\cdot)$ represents Gamma function.

To estimate the posterior distribution $q(\mu_i, \delta_i^2) = p(\mu_i, \delta_i^2 | \gamma_i, \nu_i, \alpha_i, \beta_i)$ for the i-th target image semantic representations, we factorize the distribution in the form of conjugate prior as $q(\mu_i, \delta_i^2) = q(\mu_i)q(\delta_i^2)$, which is reformulated as:

$$p(\mu_i, \delta_i^2 | \gamma_i, \nu_i, \alpha_i, \beta_i) = \frac{\beta_i^{\alpha_i}\sqrt{\nu_i}}{\Gamma(\alpha_i)\sqrt{2\pi\delta_i^2}}(\frac{1}{\delta_i^2})^{\alpha_i+1}\exp\{-\frac{2\beta_i + \nu_i(\gamma_i - \mu_i)^2}{2\delta_i^2}\}. \tag{4}$$

Given the fusion query representations and the evidential distribution from uncertainty perceptron, we then maximize the model evidence to support the observations by maximizing the likelihood of observing the semantic covariance matrix:

$$\begin{aligned}\mathcal{L}_i^{NLL} &= -\log(p(\boldsymbol{c}_{ij}|\gamma_i, \nu_i, \alpha_i, \beta_i)) \\ &= \frac{1}{2}\log(\frac{\pi}{\nu_i}) - \alpha_i \log\Omega_i + (\alpha_i + \frac{1}{2})\log((\boldsymbol{c}_{ij} - \mu_i)^2\nu_i + \Omega_i) + \log(\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + \frac{1}{2})}),\end{aligned} \tag{5}$$

where $\Omega_i = 2\beta_i(1 + \nu_i)$. Compared to directly imposing the hard-encoded labels on the distances between the fusion query and target, introducing evidential distribution to capture semantic interactions between various instances facilitates reliable similarity measurements in a more nuanced way.

Uncertainty-Guided Semantic Calibration. From the perspective of Bayesian Inference, NIG distribution is a conjugate prior of Gaussian distribution, and its corresponding parameter could be intuitively interpreted as virtual observations. To reveal the shape characteristics, the mean could be regarded as the sample mean calculated from ν virtual observations with the mean of γ, and the variance could be considered as an estimation based on α virtual observations with the mean of γ and the sum of squared deviations 2ν. Thus, the total evidence, which is the sum of all the inferred virtual observations comprised of all the virtual observation information of means and variances, is defined as $2\nu + \alpha$. Based on the model parameters of the uncertainty model in Section 3.2, the statistical moments of target semantics are computed through first-order moments of the NIG distribution:

$$\mathbb{E}[\mu] = \gamma, \quad \mathbb{E}[\delta^2] = \frac{\beta}{\alpha - 1}, \quad \text{Var}[\mu] = \frac{\beta}{\nu(\alpha - 1)}, \tag{6}$$

where the latter two notions could also be interpreted as aleatoric and epistemic uncertainty of semantic distribution. The total uncertainty is further obtained by:

$$u_i = \mathbb{E}[\delta_i^2] + \text{Var}[\mu_i] = \frac{\beta_i}{\alpha_i - 1} + \frac{\beta_i}{\nu_i(\alpha_i - 1)} = \frac{\beta_i(\nu_i + 1)}{\nu_i(\alpha_i - 1)}. \tag{7}$$

Note that semantic distribution is imbalanced in the realistic retrieving process. For instance, semantics related to colors and objects tend to appear frequently, while those involving specific details like "whimsical and vintage" are rare. Consequently, the top retrieval results tend to predict semantic categories that are more commonly represented. To avoid imbalanced learning from diverse semantics, we introduce uncertainty coefficients on cross-entropy computation based on the semantic

uncertainty fitted by NIG distribution. When a semantic embodies high uncertainty, the corresponding weight is ought to set larger, so that more penalty would be enforced on this sample during the training phase, avoiding neglecting infrequent semantic categories. The refined balanced contrastive loss is:

$$\mathcal{L}_{bc} = -\frac{1}{B}\sum_{i=1}^{B} u_i \cdot \log \frac{\exp(\boldsymbol{s}_{ii}/\tau)}{\sum_{j=1}^{B}\exp(\boldsymbol{s}_{ij}/\tau)}. \tag{8}$$

Penalty on Misleading Evidence. As the likelihood-based loss maximizes the evidence, it may lead to evidence magnification associated with the target, whereas the model discrimination on matched and hard negative samples would be limited. In particular, explicit semantics tend to overpower the outcomes and mislead the model training, whereas those with limited occurrences could be overlooked during this process. Since the misleading evidence of dominant semantics could be effective in most cases, the model is prone to maintain this incorrect evidence. It is contradictory to our target to actively reveal uncertainty when dealing with ambiguous decision boundaries instead of giving wrong predictions. Thus, we introduce the regularization term on misleading evidence to ensure robust ranking from reliable evidence. Note that $\mathbb{E}[\delta^2]$ and $\text{Var}[\mu]$ in Eq. 6 both demonstrate that uncertainty shows positive correlations with the parameter β and evidence by virtual observation theory corresponds to ν and α. Theoretical analysis is illustrated in the supplementary material. Hence, the regularization term is defined as:

$$\mathcal{L}_i^{REG} = (\mathbf{c}_{ij} - \gamma_i)^2 \cdot (2\nu_i + \alpha_i + \frac{1}{\beta_i}). \tag{9}$$

The overall uncertainty loss combines the likelihood function to maximize the model evidence and penalty for misleading evidence with controllable weight λ_1:

$$\mathcal{L}_{unc} = \frac{1}{B}\sum_{i}^{B}(\mathcal{L}_i^{NLL} + \lambda_1 \mathcal{L}_i^{REG}). \tag{10}$$

3.3 Implicit Concept Matching

Apart from uncertainty matching caused by underlying semantic covariance and distribution imbalance, complicated information sources from different modalities in the hybrid-modal query also introduce ambiguity when composing the fusion query. The reference image contains redundant visual information, *e.g.*, objects and attributes that would be replaced in the modification sentences, yet these salient features are inclined to be amplified through attention layers in the transformer-based Q-Former structure, significantly corrupting the semantic representation of the fusion query by substantial noise. Moreover, learnable prompts $\boldsymbol{p}$ involved in the computation of fusion query attempts to capture visual semantics aligned with text, however the learnable prompts $\boldsymbol{p}$ may have unclear concepts due to the misleading visual redundancy. Simply equipping supervision between fusion queries and candidates could result in substantial redundant information in embedding $\boldsymbol{p}$, especially repeatedly expressing semantics already revealed by text embeddings of modifiers.

In the training phase, we additionally integrate virtual guidance $\boldsymbol{p}^*$ to directly lead the learning of prompt and fully unleash the potentials of $\boldsymbol{p}$. MSE loss is utilized to align the learnable embeddings with virtual guidance $\boldsymbol{p}^*$ as $\mathcal{L}_{al} = \|\boldsymbol{p} - \boldsymbol{p}^*\|$. To encourage the learnable queries to acquire highly correlated messages with the targets, we deploy a symmetrical representation $\boldsymbol{q}^*$ to mirror the fusion query $\boldsymbol{q}$, as shown in Fig. 2, which combines virtual guidance $\boldsymbol{p}^*$ and text tokens of modifiers. After encoding by the text encoder using shared weights, we yield the implicit query $\boldsymbol{q}^*$, which is expected to be aligned with the target features:

$$\mathcal{L}_{ce} = -\frac{1}{B}\sum_{i=1}^{B}\log \frac{\exp(s_{ii}^*/\tau)}{\sum_{j=1}^{B}\exp(s_{ij}^*/\tau)}, \tag{11}$$

where s_{ij}^* is the cosine similarity between the implicit query $\boldsymbol{q}^*$ and target images.

In order to force the learnable embeddings to grasp implicit semantics in reference images rather than duplicated semantics mentioned in the modification text, we incorporate orthogonal loss to differentiate the retention and modification characteristics:

$$\mathcal{L}_{ort} = \sum_{i\neq j}(\text{Cov}(\boldsymbol{p}^*, \boldsymbol{m})_{ij})^2 = \sum_{i\neq j}(\frac{1}{B}((\boldsymbol{p}^*)^\top \boldsymbol{m})_{ij})^2. \tag{12}$$

The overall training objective is the aggregation of all the loss functions as: $\mathcal{L} = \mathcal{L}_{bc} + \lambda_2 \mathcal{L}_{unc} + \mathcal{L}_{ort} + \mathcal{L}_{al} + \mathcal{L}_{ce}$, where λ_2 is a trade-off parameter. The first two terms are dedicated to quantifying the uncertainty by NIG priors and calibrating the imbalanced correlations, and the remaining terms provide soft supervision from the implicit query to minimize fusion ambiguity.

Table 1: Retrieval results on FashionIQ. The best results are marked in **bold**.

Methods	Dress		Shirt		Toptee		Average		
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	Mean
VAL [18]	21.12	42.19	21.03	43.44	25.64	49.49	22.60	45.04	33.82
CIRPLANT [48]	14.38	34.66	13.64	33.56	16.44	38.34	14.82	35.52	25.17
CoSMo [49]	21.39	44.45	16.90	37.49	21.32	46.02	19.87	42.65	31.26
CLVC-Net [50]	29.85	56.47	28.75	54.76	33.50	64.00	30.70	58.41	44.56
ARTEMIS [24]	27.16	52.40	21.78	43.64	29.20	54.83	26.05	50.29	38.17
FashionVLP [30]	26.77	53.20	22.67	46.22	28.51	57.47	25.98	52.30	39.14
NSFSE [28]	31.12	55.73	24.58	45.85	31.93	58.37	29.17	53.24	41.26
CLIP4Cir [45]	31.63	56.67	36.36	58.00	38.19	62.42	35.39	59.03	47.21
CRN [51]	30.20	57.15	29.17	55.03	33.70	63.91	31.02	58.70	44.86
MGUR [4]	32.61	61.34	33.23	62.55	41.40	72.51	35.75	65.47	50.61
SPN [34]	38.82	62.92	45.83	66.44	48.80	71.29	44.48	66.88	55.68
FAME-ViL [52]	42.19	67.38	47.64	68.79	50.69	73.07	46.84	69.75	58.29
CaLa [22]	42.38	66.08	46.76	68.16	50.93	73.42	46.69	69.22	58.05
SPRC [31]	49.18	72.43	55.64	73.89	59.35	78.58	54.92	74.97	64.85
CCIN [53]	**49.38**	72.58	55.93	74.14	57.93	77.56	54.41	74.76	64.59
RUNC (Ours)	48.93	**73.53**	**57.26**	**75.32**	**60.38**	**79.86**	**55.52**	**76.23**	**65.88**

4 Experiments

4.1 Experimental Settings

Datasets. The proposed RUNC is employed on two widely-used composed image retrieval datasets covering various modification requirements in real-life retrieval scenarios. **FashionIQ** [54], concentrating on fashion item retrieval, addresses the retrieval for modifications in attributes including colors, patterns, textures, and design details across dress, toptee, and shirt categories. The whole dataset contains 77,684 fashion pictures and each matched triplet is constituted of a reference image, a modification sentence, and one target image. Following [1, 18], the dataset is split by the proportion of 3:1:1 for training, validating, and testing respectively. **CIRR** [48] involves more natural scenes and query texts more focus on alterations in the relationships among subjects, backgrounds, and multiple subjects within intricate images. It consists of 21,552 images collected from the NLVR^2 dataset [55] and constructs 36,554 matched pairs. To further evaluate the model when facing different scenarios, CIRR also provides a subset setting and each subset includes six visually similar images.

Evaluation Metrics. Following [24, 34, 31], we employ the Recall rate at K (R@K) as the main metric to evaluate the model performance, which is defined as the ratio of matched ground-truth images ranked in the top-K predictions by the model. In FashionIQ, R@10 and R@50 results are shown on dresses, toptees, and shirts. In CIRR, apart from R@1, R@5, R@10, and R@50 metrics, we also provide $\mathrm{R}_{\mathrm{subset}}$@$K$ results evaluated in the subsets.

Implementation Details. We exploited the visual and textual encoders as BLIP-$2_{\mathrm{ViT-G/14}}$ model and initialized parameters from pre-trained EVA-CLIP [32] weights. The visual encoder remained frozen and the remaining layers were fine-tuned in the training phase. The virtual guidance was disabled during the inference phase. The uncertainty perceptron was implemented as one feed-forward network (two linear layers) with a softplus activation function. The dropout rate was set as 0.2. The dimensions of fusion and candidate features were fixed as 256 in the embedding space and the number of learnable queries was set as 32. We set λ_1 as 0.01 to in Eq. 10. We used AdamW optimizer and set the learning rate as 2×10^{-5} in FashionIQ and 1×10^{-5} in CIRR with cosine annealing decay. The training and inference time of the proposed model are 214.3s and 27.4s respectively. The

Table 2: Retrieval results on CIRR test set. The best results are marked in **bold**.

Methods	Recall@K				R_{subset}@K			Avg(R@5, R_{subset}@1)
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
TIRG [1]	11.04	35.08	51.27	83.29	23.82	45.65	64.55	29.45
CIRPLANT [48]	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
ARTEMIS [24]	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
NSFSE [28]	20.70	52.50	67.96	90.74	44.20	65.53	78.50	48.35
CLIP4Cir [45]	33.59	65.35	77.35	95.21	62.39	81.81	92.02	63.87
CompoDiff [56]	22.35	54.36	73.41	91.77	35.84	56.11	76.60	45.10
BLIP4CIR [21]	40.17	71.81	83.18	95.69	72.34	88.70	95.23	72.07
SSN [57]	43.91	77.25	86.48	97.45	71.76	88.63	95.54	74.51
SPN [34]	45.33	78.07	87.61	98.17	73.93	89.28	95.61	76.00
CaLa [22]	49.11	81.21	89.59	98.00	76.27	91.04	96.46	78.74
SPRC [31]	51.96	82.12	89.74	97.69	80.65	92.31	96.60	81.39
ENCODER [20]	46.10	77.98	87.16	94.64	76.92	90.41	95.95	77.45
DIPNEC [3]	47.24	80.20	89.07	97.87	73.97	89.74	95.72	77.09
RUNC (Ours)	**53.81**	**83.47**	**91.11**	**98.22**	**80.87**	**92.36**	**96.94**	**82.17**

experiments were implemented in Pytorch on a single NVIDIA A800 GPU and trained for 30 epochs for FashionIQ and 50 epochs for CIRR[2].

4.2 Comparison with State of the Arts

Table 1 and Table 2 report quantitative comparisons of our RUNC with the advanced methods on FashionIQ and CIRR datasets, respectively. Detailed architecture descriptions of the compared methods are illustrated in the supplemental material. The proposed RUNC achieves competitive performances on benchmarks of interactive image retrieval. Specifically, a gain of 1.26% on mean R@50 in FashionIQ and a rise of 1.37% on R@10 in CIRR compared with SPRC [31] verify the effectiveness of our proposal. For fashion retrieval, the consistent growth across R@10 and R@50 metrics suggests that this approach accurately grasps user requirements and also adeptly caters to long-tail search demands. Compared with CaLa [22] and SPRC [31] in the same backbone, the overall improvements are significant, as more intrinsic semantic correlations underlying similar candidate fashion images and unclear modification descriptions in fashion queries bring uncertainty when comparing hard negatives and targets. By incorporating evidential distribution to estimate the uncertainty and calibrate the semantic imbalance, RUNC provides more reliable and robust predictions overall. Apart from recall results on all the images in the gallery, the increased metrics on subset settings as 3.95% on $\text{Recall}_{\text{subset}}$@1 in CIRR compared with the latest ENCODER [20] implicates that our proposed RUNC could identify the authentic intention from the bi-modal query when selecting from a set of closely resembling candidates.

4.3 Ablation Analysis

Analysis of Effective Components. To assess the efficacy of the model design in this work, ablative results on independent components are shown in Table 3, where "w/o UE" means disabling uncertainty estimation on semantic correlations, "w/o UGC" refers to using original form of contrastive loss without uncertainty coefficients in Eq. 8, "w/o ICM" means removing implicit concept matching, and "w/o SC" means replace the semantic covariance matrix $\boldsymbol{c}$ with ground-truth labels in Eq. 5. The remarkable drop in recall rates of "w/o UE" demonstrates that uncertainty quantification is crucial to boost the performance by perceiving the underlying correlations in candidates, and results of "w/o UGC" further verifies that uncertainty-aware calibration could avoid over-reliance on dominant categories and promote effective learning from negative samples. As the decline of results in "w/o ICM" shows, implicit query guides the distillation of inherent visual elements from reference images by aligning with the targets and independent constraints with textual embeddings. Comparing the proposed model and "w/o SC", the discrepancy highlights the significance of underlying associations in visual candidates and the effectiveness of combining evidential regression with semantic covariance in quantifying semantic ambiguity.

[2]Code is available at: RUNC-source.

Table 3: Ablative study on effective components of RUNC.

Models	FashionIQ		CIRR	
	R@10	R@50	R@5	R_{subset}@1
w/o UE	53.07	74.77	81.51	78.23
w/o UGC	53.79	75.03	82.28	78.07
w/o ICM	54.76	75.22	82.47	79.15
w/o SC	54.19	75.14	82.30	78.88
Ours	**55.52**	**76.23**	**83.47**	**80.87**

Table 4: Analysis of uncertainty estimation on R@50 metric.

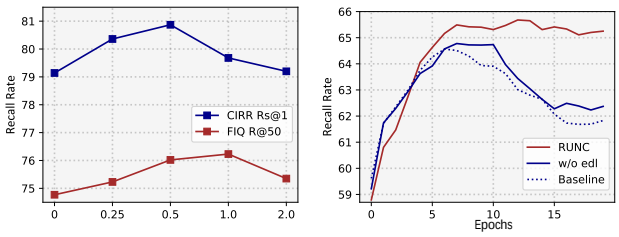
Models	Dress	Shirt	Toptee	Mean
GMM	71.69	73.26	77.41	74.12
BMM	71.83	73.99	77.91	74.58
EDC	72.83	75.22	78.74	75.60
MGUR	72.19	74.09	78.73	75.00
MPC	72.29	74.53	78.84	75.22
Ours	**73.53**	**75.32**	**79.86**	**76.23**

Analysis of Uncertainty Estimation. As shown in Table 4, we also investigate different models to quantify the uncertainty and lead to the following observations. i) "GMM" and "BMM" are sensitive to noise and fail to model aleatoric uncertainty, resulting in unstable retrieval results. Besides, the estimation dependent on the EM algorithm brings computation cost and retrieval latency. ii) Compared to Evidential Dirichlet Classification (EDC) [41], the improvement of RUNC implies that NIG distribution flexibly handles heteroscedastic noise, which better aligns with the coupling semantic correlations in this retrieval task. iii) Though MGUR [4] and MPC [58] utilize probabilistic distribution to model the data uncertainty, insufficient semantic supervision may result in additional noise by nondirective distributions. In comparison, high-order priors in RUNC is superior in quantifying the underlying uncertainty and enhance the interpretability of ranking results.

4.4 Further Discussion

Impact of Uncertainty Supervision. To evaluate the sensitivity of λ_2 controlling the uncertainty estimation loss in different datasets, Fig. 3(a) presents recall rates on different settings. The optimal value of λ_2 for FashionIQ is slightly bigger than CIRR for more severe aleatoric uncertainty issues in FashionIQ. More sensitivity analysis could be referred to in the supplementary material.

Impact of Evidential Learning. To verify the necessity of introducing evidential learning to enhance the robustness, we also conduct an ablative setting as directly using semantic covariance matrix $\boldsymbol{c}$ as weights in contrastive learning, denoted as "w/o edl" in Fig. 3(b). Although correlation weights could promote discriminant learning compared with baseline, it is less superior than proposed evidential learning for accumulating errors from precomputed similarities and neglecting the aleatoric uncertainty by data noise and false negatives.



(a) Param λ_2 (b) Evidential Learning

Figure 3: Analysis of λ_2 and evidential learning.

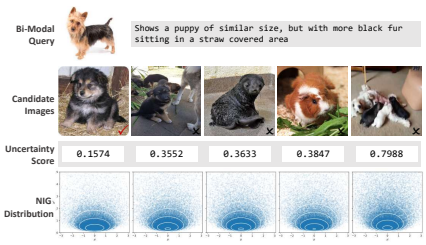


Figure 4: Visualization of uncertainty on CIRR.

4.5 Visualization Analysis

Visualization of Uncertainty. To enhance comprehension of the retrieval results based on uncertainty, we also present retrieval examples to visualize the uncertainty quantification in Fig. 4. For the rightmost image with low resolution and sharing semantics like "black fur" and "puppy" with the user intent, the estimated uncertainty value leads to an increase in the penalty by Eq. 8.

Visualization of Implicit Concept Learning. Additionally, we also present activation heatmaps by GradCAM [59] on the last module in the vision encoder with other top-ranking results in Figure 5. The prompt embedding $\boldsymbol{p}$ highlights the area corresponding to the dog which is coherent with the reference images, and $\boldsymbol{m}$ underlines the area of the yellow pillow on the couch. The collaboration between $\boldsymbol{p}$ and $\boldsymbol{m}$ enables the model to accurately select the ground-truth image from other candidates, especially enhancing the perception of object contours, categories, or backgrounds provided by the reference image in $\boldsymbol{p}$.

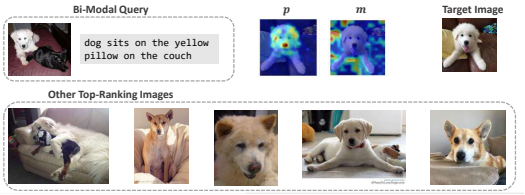


Figure 5: Heatmaps on Target Images with Other Top ranking results.

Failure Cases. Failure cases of the proposed RUNC are shown in Figure 6, where the ground-truth images (red boxes) are not ranked the top by our model. For the first example, since the query text in the first instance does not specify the need to change the type of animal in the image, the top-1 image in this model depicts a monkey sleeping on the bend of the tree trunk, which can be regarded as a false negative sample. It also indicates that the proposed RUNC is capable of perceiving the basic query requirements of users effectively. Regarding the second example, due to the complexity of modifying the semantics in the text, and being influenced by factors such as lighting and angles, the model fails to precisely grasp the requirements of "hind legs" and "industrial setting" during measuring the distance between the query and candidate images.

5 Conclusion

In this paper, we introduced a novel robust uncertainty calibration model dubbed RUNC to quantify the uncertainty and mitigate the imbalanced semantic learning for interactive image retrieval. To encourage learning from infrequent semantic concepts and mitigate interference caused by semantic overlaps, high-order evidential priors are deployed to estimate the aleatoric and epistemic uncertainty, and target distribution is further adjusted based on uncertainty coefficients. Moreover, we employed an orthogonal loss between explicit textual embeddings and implicit queries to minimize the semantic ambiguity of fusion query from reference images and modification texts. The effectiveness and robustness of RUNC have been demonstrated by extensive experimental results and ablation studies. The potential for extending this approach to other multi-modal learning tasks provides promising prospects for further investigation.

Limitations. Since the proposed RUNC utilizes evidential learning that relies on the hypothesis of prior distribution, it may face challenges in continuous learning and lack adaptability to dynamic environments. When the data distribution rapidly changes, frequent updates of uncertain distribution parameters are required, making it difficult for the model to quickly adjust confidence estimates. Besides, a certain amount of data is required in RUNC to fit the prior distribution. In situations involving small sample sizes, the model may struggle to accurately learn the parameters of uncertainty, potentially resulting in either overestimation or underestimation of confidence levels.

Figure 6: Failure cases in the proposed RUNC.

6 Acknowledgments

This work is supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012, KJZD20230923114916032, GJHZ20240218113604008).

References

[1] N. Vo, L. Jiang, C. Sun, K. Murphy, L. Li, L. Fei-Fei, and J. Hays, "Composing text and image for image retrieval - an empirical odyssey," in *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, 2019, pp. 6439–6448.

[2] Y. Chen and L. Bazzani, "Learning joint visual semantic matching embeddings for language-guided retrieval," in *European Conference on Computer Vision*, vol. 12367. Springer, 2020, pp. 136–152.

[3] Y. Wang, W. Huang, and C. Yuan, "Aligning composed query with image via discriminative perception from negative correspondences," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2025, pp. 8078–8086.

[4] Y. Chen, Z. Zheng, W. Ji, L. Qu, and T. Chua, "Composed image retrieval with text feedback via multi-grained uncertainty regularization," in *International Conference on Learning Representations*. OpenReview.net, 2024.

[5] Z. Sun, D. Jing, G. Yang, N. Fei, and Z. Lu, "Leveraging large vision-language model as user intent-aware encoder for composed image retrieval," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2025, pp. 7149–7157.

[6] Y. Lin, Y. Xie, D. Chen, Y. Xu, C. Zhu, and L. Yuan, "REVIVE: regional visual representation matters in knowledge-based visual question answering," in *Annual Conference on Neural Information Processing Systems*, 2022.

[7] Y. Peng, C. Hao, X. Hu, J. Peng, X. Geng, and X. Yang, "LIVE: learnable in-context vector for visual question answering," in *Annual Conference on Neural Information Processing Systems*, 2024.

[8] W. Lin, J. Chen, J. Mei, A. Coca, and B. Byrne, "Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering," in *Annual Conference on Neural Information Processing Systems*, 2023.

[9] T. Gupta and A. Kembhavi, "Visual programming: Compositional visual reasoning without training," in *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, 2023, pp. 14 953–14 962.

[10] H. Al-Tahan, Q. Garrido, R. Balestriero, D. Bouchacourt, C. Hazirbas, and M. Ibrahim, "Unibench: Visual reasoning requires rethinking vision-language beyond scaling," in *Annual Conference on Neural Information Processing Systems*, 2024.

[11] S. Jaiswal, D. Roy, B. Fernando, and C. Tan, "Learning to reason iteratively and parallelly for complex visual reasoning scenarios," in *Annual Conference on Neural Information Processing Systems*, 2024.

[12] X. Jiang, Z. Huang, X. Xu, J. Song, F. Shen, and H. T. Shen, "Phgc: Procedural heterogeneous graph completion for natural language task verification in egocentric videos," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8615–8624.

[13] F. Zhang, M. Xu, and C. Xu, "Geometry sensitive cross-modal reasoning for composed query based image retrieval," *IEEE Trans. Image Process.*, vol. 31, pp. 1000–1011, 2022.

[14] F. Huang, L. Zhang, X. Fu, and S. Song, "Dynamic weighted combiner for mixed-modal image retrieval," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2024, pp. 2303–2311.

[15] Y. Chen, J. Zhou, and Y. Peng, "SPIRIT: style-guided patch interaction for fashion image retrieval with text feedback," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 20, no. 6, pp. 167:1–167:17, 2024.

[16] Y. K. Jang, D. Kim, Z. Meng, D. Huynh, and S. Lim, "Visual delta generator with large multi-modal models for semi-supervised composed image retrieval," in *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, 2024, pp. 16 805–16 814.

[17] H. Wen, X. Song, J. Yin, J. Wu, W. Guan, and L. Nie, "Self-training boosted multi-factor matching network for composed image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3665–3678, 2024.

[18] Y. Chen, S. Gong, and L. Bazzani, "Image search with text feedback by visiolinguistic attention learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, 2020, pp. 2998–3008.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, vol. 139. PMLR, 2021, pp. 8748–8763.

[20] Z. Li, Z. Chen, H. Wen, Z. Fu, Y. Hu, and W. Guan, "ENCODER: entity mining and modification relation binding for composed image retrieval," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2025, pp. 5101–5109.

[21] Z. Liu, W. Sun, Y. Hong, D. Teney, and S. Gould, "Bi-directional training for composed image retrieval via text prompt learning," in *Winter Conference on Applications of Computer Vision*. IEEE, 2024, pp. 5741–5750.

[22] X. Jiang, Y. Wang, M. Li, Y. Wu, B. Hu, and X. Qian, "Cala: Complementary association learning for augmenting comoposed image retrieval," in *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2024, pp. 2177–2187.

[23] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, "Composed image retrieval using contrastive learning and task-oriented clip-based features," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 3, 2023.

[24] G. Delmas, R. S. de Rezende, G. Csurka, and D. Larlus, "ARTEMIS: attention-based retrieval with text-explicit matching and implicit similarity," in *International Conference on Learning Representations*. OpenReview.net, 2022.

[25] E. Dodds, J. Culpepper, S. Herdade, Y. Zhang, and K. Boakye, "Modality-agnostic attention fusion for visual search with text feedback," *CoRR*, vol. abs/2007.00145, 2020.

[26] B. Schroeder and S. Tripathi, "Structured query-based image retrieval using scene graphs," in *Proceedings of the Computer Vision and Pattern Recognition Conference*. Computer Vision Foundation / IEEE, 2020, pp. 680–684.

[27] M. U. Anwaar, E. Labintcev, and M. Kleinsteuber, "Compositional learning of image-text query for image retrieval," in *Winter Conference on Applications of Computer Vision*. IEEE, 2021, pp. 1139–1148.

[28] Y. Wang, L. Liu, C. Yuan, M. Li, and J. Liu, "Negative-sensitive framework with semantic enhancement for composed image retrieval," *IEEE Trans. Multim.*, vol. 26, pp. 7608–7621, 2024.

[29] M. Shin, Y. Cho, B. Ko, and G. Gu, "Rtic: Residual learning for text and image composition using graph convolutional network," *CoRR*, vol. abs/2104.03015, 2021.

[30] S. Goenka, Z. Zheng, A. Jaiswal, R. Chada, Y. Wu, V. Hedau, and P. Natarajan, "Fashionvlp: Vision language transformer for fashion retrieval with feedback," in *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, 2022, pp. 14 085–14 095.

[31] Y. Bai, X. Xu, Y. Liu, S. Khan, F. Khan, W. Zuo, R. S. M. Goh, and C. Feng, "Sentence-level prompts benefit composed image retrieval," in *International Conference on Learning Representations*. OpenReview.net, 2024.

[32] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning*, vol. 202. PMLR, 2023, pp. 19 730–19 742.

[33] J. Kim, Y. Yu, H. Kim, and G. Kim, "Dual compositional learning in interactive image retrieval," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2021, pp. 1771–1779.

[34] Z. Feng, R. Zhang, and Z. Nie, "Improving composed image retrieval via contrastive learning with scaling positives and negatives," in *ACM International Conference on Multimedia*. ACM, 2024, pp. 1–10.

[35] Y. Liu, R. Shen, and X. Shen, "Novel uncertainty quantification through perturbation-assisted sample synthesis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 7813–7824, 2024.

[36] X. Peng, F. Qiao, and L. Zhao, "Out-of-domain generalization from a single source: An uncertainty quantification approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1775–1787, 2024.

[37] G. Franchi, A. Bursuc, E. Aldea, S. Dubuisson, and I. Bloch, "Encoding the latent posterior of bayesian neural networks for uncertainty quantification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2027–2040, 2024.

[38] Y. Cui, W. Yao, Q. Li, A. B. Chan, and C. J. Xue, "Accelerating monte carlo bayesian prediction via approximating predictive uncertainty over the simplex," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 4, pp. 1492–1506, 2022.

[39] A. Zhou and S. Levine, "Amortized conditional normalized maximum likelihood: Reliable out of distribution uncertainty estimation," in *International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 12 803–12 812.

[40] F. Lu, K. Zhu, W. Zhai, K. Zheng, and Y. Cao, "Uncertainty-aware optimal transport for semantically coherent out-of-distribution detection," in *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, 2023, pp. 3282–3291.

[41] M. Sensoy, L. M. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Annual Conference on Neural Information Processing Systems*, 2018, pp. 3183–3193.

[42] X. Jiang, X. Xu, L. Zhu, Z. Sun, A. Cichocki, and H. T. Shen, "Resisting noise in pseudo labels: Audible video event parsing with evidential learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 6, pp. 10 874–10 888, 2025.

[43] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," in *Annual Conference on Neural Information Processing Systems*, 2020.

[44] J. Lou, W. Liu, Z. Chen, F. Liu, and J. Cheng, "Elfnet: Evidential local-global fusion for stereo matching," in *International Conference on Computer Vision*. IEEE, 2023, pp. 17 738–17 747.

[45] A. Baldrati, M. Bertini, T. Uricchio, and A. D. Bimbo, "Conditioned and composed image retrieval combining and partially fine-tuning clip-based features," in *CVPR Workshops*. IEEE, 2022, pp. 4955–4964.

[46] A. Malinin and M. J. F. Gales, "Predictive uncertainty estimation via prior networks," in *Annual Conference on Neural Information Processing Systems*, 2018, pp. 7047–7058.

[47] M. Sensoy, L. M. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Annual Conference on Neural Information Processing Systems*, 2018, pp. 3183–3193.

[48] Z. Liu, C. R. Opazo, D. Teney, and S. Gould, "Image retrieval on real-life images with pre-trained vision-and-language models," in *International Conference on Computer Vision*. IEEE, 2021, pp. 2105–2114.

[49] S. Lee, D. Kim, and B. Han, "Cosmo: Content-style modulation for image retrieval with text feedback," in *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, 2021, pp. 802–812.

[50] H. Wen, X. Song, X. Yang, Y. Zhan, and L. Nie, "Comprehensive linguistic-visual composition network for image retrieval," in *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2021, pp. 1369–1378.

[51] Q. Yang, M. Ye, Z. Cai, K. Su, and B. Du, "Composed image retrieval via cross relation network with hierarchical aggregation transformer," *IEEE Transactions on Image Processing*, vol. 32, pp. 4543–4554, 2023.

[52] X. Han, X. Zhu, L. Yu, L. Zhang, Y. Song, and T. Xiang, "Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks," in *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, 2023, pp. 2669–2680.

[53] L. Tian, J. Zhao, Z. Hu, Z. Yang, H. Li, L. Jin, Z. Wang, and X. Li, "CCIN: compositional conflict identification and neutralization for composed image retrieval," in *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, 2025, pp. 3974–3983.

[54] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, "Fashion IQ: A new dataset towards retrieving images by natural language feedback," in *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, 2021, pp. 11 307–11 317.

[55] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, "A corpus for reasoning about natural language grounded in photographs," in *ACL*. Association for Computational Linguistics, 2019, pp. 6418–6428.

[56] G. Gu, S. Chun, W. Kim, H. Jun, Y. Kang, and S. Yun, "Compodiff: Versatile composed image retrieval with latent diffusion," *Trans. Mach. Learn. Res.*, vol. 2024, 2024.

[57] X. Yang, D. Liu, H. Zhang, Y. Luo, C. Wang, and J. Zhang, "Decompose semantic shifts for composed image retrieval," *CoRR*, vol. abs/2309.09531, 2023.

[58] N. Andrei, Y. Chen, and Z. Akata, "Probabilistic compositional embeddings for multimodal image retrieval," in *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, 2022, pp. 4546–4556.

[59] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *International Conference on Computer Vision*. IEEE, 2017, pp. 618–626.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations of the work performed by the authors in the Section "Limitations".

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Our paper only involves a small amount of theory, which has been illustrated in the Methodology (as seen in Section 3) and supplementary material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We disclosed all the information needed to reproduce the main experimental results of the paper. We also provided implementation details in the Experiments (as seen in Section 4.1).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The manuscript provides detailed information to reproduce the results. We also provided the corresponding references and links.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The manuscript provides detailed information about the experimental setting and implementation details in Experiments (as seen in Section 4.1).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Error bar is not commonly used in this Composed Image Retrieval task.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We presented our compute resources in Section 4.1

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed both potential positive societal impacts and negative societal impacts of the work performed in this paper and supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets (e.g., code, data, models) used in the paper are properly credited, and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is used only for polishing the writing of this manuscript and does not impact the core methodology, scientific rigorousness, or originality of the research

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.