ROBUST UNSUPERVISED CROSS-MODAL HASHING VIA CAUSAL SUBJECT-CONTEXT INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In unsupervised cross-modal hashing, similarity relations are used to guide hash learning and thus directly shape the learned Hamming space. Although existing methods seek to obtain more reliable similarity relations through various estimation and refinement strategies, they generally derive such relations from representations where retrieval-relevant subject semantics may still be entangled with contextual factors, such as shared backgrounds, textual styles, or incidental co-occurrences. Consequently, context-induced associations may be encoded as semantic similarities and propagated through iterative similarity updates, thereby misleading subsequent hash learning. To address this issue, we propose Causal Subject-Context Intervention Hashing (CSCIH). Specifically, Causal Subject-Context Intervention (CSCI) learns subject and context oriented representations and performs an approximate differentiable intervention by restricting residual refinement to context-oriented dimensions while preserving subject oriented ones. It further evaluates whether subject information remains consistent across alternative contextual states and summarizes this evidence as a subject consistency score. Moreover, CSCIH derives complementary similarity relations from subject oriented representations and continuous hash representations. Health-Aware Structure Refresh (HASR) adaptively balances these relations according to subject consistency and training progress, enabling more reliable similarity updates throughout training. Finally, Geometry-Aware Joint Hash Learning (GJHL) captures modality-local and cross-modal relational structures in both Euclidean and hyperbolic spaces and integrates them into a continuous teacher representation to guide modality-specific hash learning. Experiments on three benchmark datasets demonstrate that CSCIH consistently outperforms representative unsupervised cross-modal hashing methods.

1 INTRODUCTION

The rapid growth of multimodal data has made efficient cross-modal retrieval increasingly important. Cross-modal hashing maps heterogeneous modalities into a shared Hamming space, where compact binary codes enable fast retrieval with low storage cost. However, acquiring large-scale semantic annotations is often prohibitively expensive, which has motivated increasing interest in unsupervised cross-modal hashing. In the absence of category labels, these methods typically derive training similarities from the data itself or from learned representations. Consequently, reliable similarity estimation is crucial, as inaccurate estimates can misguide training and distort the learned Hamming space.

Existing methods mainly improve similarity estimation through joint-semantics affinity modeling (Su et al., 2019), correlation graphs (Zhang et al., 2018) and graph-neighbor coherence (Yu et al., 2021), community similarity quantization (Fan & Cao, 2025), or contrastive learning with cross-modal ranking (Hu et al., 2023). However, the similarities used to guide training are typically estimated from the model's evolving representations, creating a dependency between similarity estimation and representation learning. An inaccurate similarity matrix at one update can misguide representation learning in subsequent epochs, and these affected representations are then used to construct the similarity matrix at the next update. As a result, estimation errors may propagate across successive updates and progressively affect the learned Hamming space. This observation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

motivates us to consider similarity estimation across successive updates, rather than treating each update as an isolated estimation step.

Within this iterative process, a particular source of estimation error is the entanglement of retrieval-relevant subject semantics with contextual factors in multimodal representations, such as shared backgrounds, stylistic patterns, and incidental co-occurrences. These contextual cues may induce spurious associations that are interpreted as semantic similarities during similarity estimation. Once incorporated into the similarity matrix, such associations can influence subsequent representation learning and persist across later similarity updates. Existing similarity-refinement methods mainly operate on similarity relations derived from the current representation space, without explicitly addressing the subject-context entanglement in the representations from which these relations are constructed.

To address this issue, we propose Causal Subject-Context Intervention Hashing (CSCIH). Specifically, in Causal Subject-Context Intervention (CSCI), subject and context oriented representations are learned, and an approximate differentiable intervention is achieved by applying residual refinement only to context-oriented dimensions while keeping subject oriented ones unchanged. The consistency of subject information across alternative contextual states is then evaluated to obtain a subject consistency score. In addition, CSCIH constructs complementary similarity relations using subject oriented representations and continuous hash representations. Health-Aware Structure Refresh (HASR) dynamically combines these relations according to subject consistency and training progress, while using the fixed initial similarity matrix as an anchor to limit accumulated drift. These designs support more reliable similarity updates throughout training without directly replacing the relations learned by the retrieval functions. Finally, Geometry-Aware Joint Hash Learning (GJHL) preserves modality-local relational structures and represents cross-modal relations in complementary Euclidean and hyperbolic spaces. The resulting relational information is then incorporated into a continuous teacher representation to guide modality-specific hash learning.

The main contributions of this work can be summarized as follows:

- We propose a subject-context intervention strategy with consistency-guided structure refresh for similarity updates. CSCI learns subject- and context-oriented representations, performs an approximate differentiable intervention, and assesses subject consistency across alternative contextual states. HASR then balances complementary similarity relations according to subject consistency and training progress, while the initial similarity matrix constrains drift.
- We introduce a geometry-aware joint hash learning mechanism that preserves modality-local relational structures and represents cross-modal relations in complementary Euclidean and hyperbolic spaces. It further integrates modality-local and joint relational information into a continuous teacher representation to guide modality-specific hash learning under the current similarity matrix during training.
- Extensive experiments on three benchmark datasets demonstrate that CSCIH consistently outperforms representative unsupervised cross-modal hashing methods.

2 Related Work

2.1 Unsupervised Cross-Modal Hashing

Unsupervised cross-modal hashing aims to learn binary codes from paired multimodal data without category labels. Existing methods mainly focus on constructing and strengthening pseudo-supervision, which can be roughly grouped into three categories(Wang et al., 2024a; Kuai et al., 2025). The first category emphasizes cross-modal alignment and reconstruction, aiming to mitigate modal gaps and establish a unified shared semantic space, including adversarial distribution matching (Zhang et al., 2018), common-space constraints (Yang et al., 2020), and the recovery of joint semantics and cross-modal correlations (Su et al., 2019; Zhu et al., 2023). The second category focuses on neighborhood structure or affinity relation modeling, such as correlation graph construction (Chen et al., 2020), joint-modal distribution modeling (Liu et al., 2020), graph-neighbor coherence preservation (Yu et al., 2021), global cross-modal ranking (Hu et al., 2023), and cross-domain semantic neighborhood transfer (Li et al., 2024a). The third category strengthens supervision by fur-

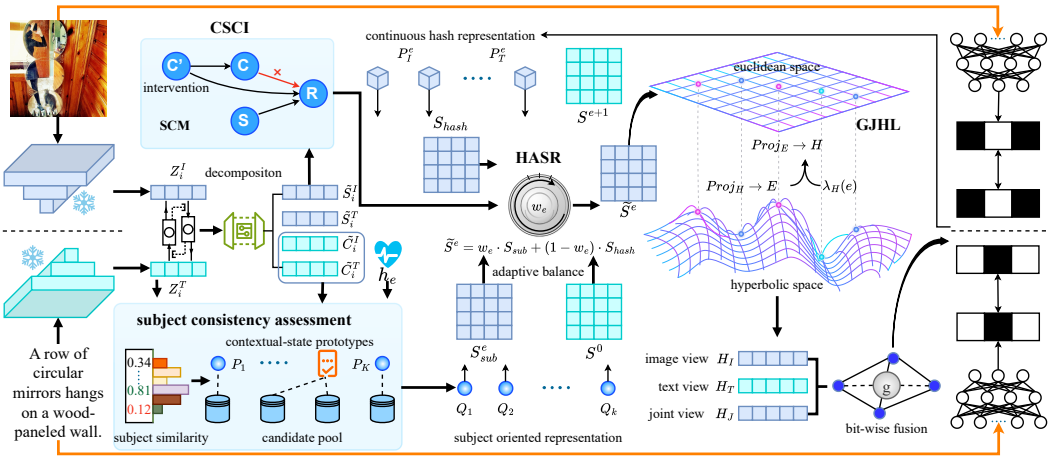


Figure 1: Overview of our proposed CSCIH. CSCI learns subject and contextoriented representations, performs an approximate differentiable intervention, and evaluates subject consistency across alternative contextual states. CSCIH then derives complementary similarity relations from subject oriented representations and continuous hash representations, which HASR adaptively balances according to subject consistency and training progress. GJHL further captures modality-local and cross-modal relational structures in Euclidean and hyperbolic spaces and integrates them into a continuous teacher representation to guide modality-specific hash learning.

ther introducing external semantic or visual guidance, such as community-level quantization (Song et al., 2025; Fan & Cao, 2025; Chen et al., 2026b), semantic self-learning (Chen et al., 2026c), and Transformer-based contrastive modeling (Yang et al., 2026b), thereby improving the quality and discriminability of pseudo-supervision. However, existing methods generally construct pseudo-relations without distinguishing subject semantics from contextual factors, making the resulting supervision susceptible to contextual interference.

2.2 Causal and Disentangled Representation Learning

Causal representation learning aims to discover latent factors that remain stable under interventions and distribution shifts (Schölkopf et al., 2021), while disentangled representation learning seeks to separate task-relevant factors from nuisance variations (Wang et al., 2024b). Combining these two perspectives, methods such as CausalVAE incorporate structural causal models into representation learning (Yang et al., 2021a). In contrast, counterfactual intervention approaches identify stable factors by examining representation changes under controlled interventions (Li et al., 2024b). These advances provide a principled basis for distinguishing intrinsic semantics from spurious contextual correlations. Recent studies have extended these ideas to multimodal representation learning. Several studies assess multimodal relations through causal sufficiency, necessity, or cross-modal intervention (Wang et al., 2025; Xu et al., 2026), and others identify intervention-stable dependencies or control confounding effects at the prototype level (Wang et al., 2026; Li et al., 2026). Causal principles have also been applied to multimodal generalization (Liang et al., 2026), visual invariance in composed image retrieval (Chen et al., 2026d), and semantic recovery in incomplete cross-modal retrieval (Chen et al., 2026a). Collectively, these studies provide a useful foundation for modeling cross-modal relations that remain stable under contextual variation.

3 Preliminary

Given an unlabeled paired image-text dataset $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$, where $x_i \in \mathbb{R}^{d_I}$ and $y_i \in \mathbb{R}^{d_T}$ denote the image and text features of the ith pair, respectively, our goal is to learn two modality-specific hash functions $f_I : \mathbb{R}^{d_I} \to [-1, 1]^c$ and $f_T : \mathbb{R}^{d_T} \to [-1, 1]^c$, where c denotes the hash-code length. For modality $m \in I, T$, we write $z_i^I = x_i$ and $z_i^T = y_i$. The ℓ_2 normalization is defined as

$\mathcal{N}(v) = v/\max(|v|_2, \epsilon)$ with $\epsilon > 0$, and is applied row-wise to matrices. The corresponding binary codes are $b_i^I = \text{sign}(f_I(\mathcal{N}(x_i)))$ and $b_i^T = \text{sign}(f_T(\mathcal{N}(y_i)))$. Paired observations are available during training to estimate the similarity relations used to guide hash learning, whereas retrieval is performed independently with the corresponding modality-specific hash function.

At epoch e, training is performed over mini-batches $\mathcal{B} \subset \mathcal{D}$ under the current similarity matrix S^e, with $S^e_{\mathcal{B}}$ denoting its mini-batch submatrix. The matrix remains fixed throughout the epoch and is revised only at scheduled similarity-matrix updates. At these updates, the representations obtained after epoch e are used to estimate S^{e+1} for the following epochs, while the current matrix is retained between updates. In this way, a newly estimated similarity matrix affects subsequent optimization rather than the epoch from which it was obtained.

4 METHODOLOGY

Existing unsupervised cross-modal hashing methods improve similarity estimation through graph-neighbor coherence, similarity refinement, and relation reweighting (Yu et al., 2021; Yang et al., 2026a). However, these strategies still operate on relations derived from learned representations. When retrieval-relevant subject semantics remain entangled with contextual factors, context-induced associations may be encoded into similarity relations and persist across subsequent similarity updates. This motivates examining subject oriented information under contextual variation rather than relying only on further relation refinement.

Causal intervention provides a natural perspective for reducing contextual confounding (Yang et al., 2021b; Liu et al., 2022). To address this issue, we propose Causal Subject-Context Intervention Hashing (CSCIH). As illustrated in Fig. 1, CSCI learns subject and context oriented representations and assesses subject consistency; HASR balances complementary similarity relations according to subject consistency and training progress; and GJHL organizes modality-local and cross-modal relational structures for hash learning.

4.1 CAUSAL SUBJECT-CONTEXT INTERVENTION

Subject semantics and contextual factors may remain entangled in multimodal representations. CSCI learns a sample-dependent gate g_i^m and derives subject and context oriented representations as

$$s_i^m = E_s^m(g_i^m \odot z_i^m), \quad c_i^m = E_c^m((1 - g_i^m) \odot z_i^m). \tag{1}$$

Here, E_s^m and E_c^m are modality-specific mappings for the two representations. Paired image and text representations are further fused into shared representations s_i and c_i, with cross-modal consistency imposed on the subject oriented representations and calibration regularization used to stabilize the gating. The modality-specific s_i^m are later used to derive similarity relations, while s_i and c_i support subject consistency assessment under contextual variation.

CSCI performs an approximate differentiable intervention by restricting residual refinement to context-oriented dimensions while preserving subject oriented ones:

$$\bar{z}_i^m = z_i^m + \eta_e(1 - g_i^m) \odot \tanh(D_m(c_i)). \tag{2}$$

Here, $\bar{z}_i^m$ is the refined representation, $\eta_e \geq 0$ controls the refinement strength, and D_m maps c_i back to the representation space of modality m. The residual is computed from the current sample's own context-oriented representation and modulated by $(1 - g_i^m)$, thereby concentrating refinement on context-oriented dimensions while largely preserving subject oriented ones.

Because these representations are learned without category labels, CSCI assesses whether their subject information remains consistent across alternative contextual states. We represent K latent contextual states by prototypes $\{p_k\}_{k=1}^K$ and assign c_i according to

$$\pi_{ik} = \frac{\exp\left(\mathcal{N}(c_i)^\top \mathcal{N}(p_k)/\tau_{\text{env}}\right)}{\sum_{\ell=1}^K \exp\left(\mathcal{N}(c_i)^\top \mathcal{N}(p_\ell)/\tau_{\text{env}}\right)}. \tag{3}$$

Here, $\tau_{\text{env}} > 0$ is the assignment temperature and $\pi_i = (\pi_{i1}, \ldots, \pi_{iK})$ is the contextual-state distribution of sample i. CSCI identifies alternative samples with similar subject oriented representations

but different contextual states. These alternatives are used only for subject consistency assessment and not for residual refinement; the selection procedure is given in the Appendix B.1.

For a mini-batch B, let r_B denote the fraction of samples with at least one such alternative, and let $H(\pi_i)$ denote the Shannon entropy of π_i, measuring uncertainty in contextual-state assignment. We combine alternative availability with assignment confidence as

$$h_B = \lambda_h r_B + (1-\lambda_h)\left(1 - \frac{1}{|B|\log K}\sum_{i\in B} H(\pi_i)\right), \quad h_e = \frac{1}{|\mathcal{B}_e|}\sum_{B\in\mathcal{B}_e} h_B. \tag{4}$$

Here, $\lambda_h \in [0,1]$ balances the two terms, $\log K$ normalizes the entropy over K states, and $\mathcal{B}_e$ is the set of mini-batches at epoch e. The resulting h_e is the subject consistency score used by HASR to regulate the contribution of similarity relations derived from the subject oriented representations.

4.2 Health-Aware Structure Refresh

CSCI provides modality-specific subject oriented representations together with the subject consistency score h_e. At a scheduled similarity-matrix update following epoch e, let P_I^e and P_T^e collect the current continuous hash representations over the training set, and let $U_s^I = [s_1^I, \ldots, s_N^I]^\top$ and $U_s^T = [s_1^T, \ldots, s_N^T]^\top$ collect the corresponding modality-specific subject oriented representations obtained by CSCI. Using a shared relation constructor $\mathcal{K}(\cdot,\cdot)$, CSCIH constructs two similarity matrices encoding complementary relations:

$$S^e_{\text{hash}} = \mathcal{K}(P_I^e, P_T^e), \quad S^e_{\text{sub}} = \mathcal{K}(U_s^I, U_s^T). \tag{5}$$

Here, S^e_{hash} reflects the relations currently encoded by the continuous hash representations, whereas S^e_{sub} captures subject oriented relations that complement the current hash-derived relations.

Because both relation sources are derived from learned representations that evolve during training, successive similarity updates based solely on these evolving relations may lead to accumulated drift. CSCIH therefore retains the initial similarity matrix S^0 as a fixed anchor. HASR adaptively balances the complementary similarity relations according to subject consistency and training progress. Let $\alpha_e \in [0,1]$ denote the training-progress schedule, $\lambda_s \geq 0$ the subject-relation scale, and $[u]_{[0,1]}$ clipping to $[0,1]$. We define

$$\omega_e = [\lambda_s \alpha_e h_e]_{[0,1]}, \quad \widetilde{S}^e = (1-\omega_e)S^e_{\text{hash}} + \omega_e S^e_{\text{sub}}, \quad S^{e+1} = \mathcal{T}\left(\lambda_0 S^0 + (1-\lambda_0)\widetilde{S}^e\right), \tag{6}$$

where $\lambda_0 \in [0,1]$ is the anchoring coefficient. The operator $\mathcal{T}(\cdot)$ projects the blended estimate back to the similarity-matrix form used for subsequent training by enforcing symmetry, self-similarity, and bounded similarity values; further implementation details are provided in the Appendix B.2. The balance therefore depends jointly on subject consistency through h_e and training progress through α_e, while S^0 constrains accumulated drift across successive similarity updates.

Similarity-matrix updates are performed only at scheduled epochs. If no update is triggered, the current similarity matrix remains unchanged, i.e., $S^{e+1} = S^e$. When an update is scheduled following epoch e, S^{e+1} is constructed only after completing that epoch using the representations and subject consistency score obtained up to that point, and is used only in subsequent optimization epochs. The resulting similarity matrix then provides the cross-modal relational constraints for GJHL, which preserves modality-local relational structures while organizing cross-modal relations through complementary Euclidean and hyperbolic geometries.

4.3 Geometry-Aware Joint Hash Learning

The current similarity matrix S^e provides cross-modal relational constraints, but these constraints do not by themselves preserve the modality-specific local continuity of image and text representations. GJHL therefore models modality-local and cross-modal relational structures separately. For modality m, sample i and its two local neighbors are aggregated as

$$\tilde{z}_i^m = \sum_{j=0}^{2} w_{ij}^m \bar{z}_{n_j(i)}^m, \tag{7}$$

where $n_0(i) = i$, $w_{ij}^m \geq 0$, and $\sum_j w_{ij}^m = 1$. Modality-specific encoding and local propagation then produce V_I and V_T, preserving intra-modal continuity without forcing the two modalities to share identical local neighborhoods.

Cross-modal relations are organized through complementary Euclidean and hyperbolic geometries. Euclidean propagation produces the base joint representation V_J^E, while a hyperbolic branch processes the same relational structure in Poincaré space to obtain V_J^H. Hyperbolic geometry has been explored for multimodal representation learning (Ramasinghe et al., 2024), and recent work suggests that Euclidean and hyperbolic geometries can provide complementary relational biases (Na et al., 2026). Analyses of hyperbolic representations further motivate treating them as a geometric inductive bias rather than assuming explicit recovery of latent hierarchy (Yang et al., 2023). We combine the two geometries as

$$V_J = V_J^E + \lambda_H(e)\left(V_J^H - V_J^E\right), \quad \lambda_H(e) \in [0,1]. \tag{8}$$

The Euclidean representation remains the base, while $\lambda_H(e)$ controls the contribution of the complementary hyperbolic structure according to training progress.

The modality-local representations V_I, V_T and the joint representation V_J are mapped to continuous representations H_I, H_T, H_J, respectively. A bit-wise fusion operator $\mathcal{F}_H$ integrates them into the continuous teacher representation

$$H = \mathcal{F}_H(H_I, H_J, H_T), \quad H \in [-1,1]^{|\mathcal{B}| \times c}. \tag{9}$$

The resulting teacher representation combines modality-local relational structures with cross-modal relations organized through complementary Euclidean and hyperbolic geometries.

To preserve local fidelity, stabilize the learned CSCI gating, and encourage binary compatibility, we optimize the continuous teacher representation using the neighborhood reconstruction objective $\mathcal{L}_{\mathrm{NBR}}$, the CSCI calibration regularizer $\mathcal{L}_{\mathrm{cal}}$, and the quantization objective $\mathcal{L}_{\mathrm{QUANT}}$. The teacher-side objective is

$$\mathcal{L}_{\mathrm{teacher}} = \lambda_{\mathrm{NBR}}\mathcal{L}_{\mathrm{NBR}} + \mathcal{L}_{\mathrm{cal}} + \lambda_{\mathrm{QUANT}}\mathcal{L}_{\mathrm{QUANT}}, \tag{10}$$

with $\lambda_{\mathrm{NBR}}, \lambda_{\mathrm{QUANT}} \geq 0$. Detailed constructions of the neighborhood reconstruction and quantization objectives are provided in the Appendix B.3.

During each epoch, GJHL is optimized under the current similarity matrix S^e to obtain the continuous teacher representation H, which then guides modality-specific hash learning.

4.4 Teacher-Guided Modality-Specific Hash Learning

Given the continuous teacher representation H from GJHL, the modality-specific hash functions f_I and f_T are optimized while retaining independent modality inputs. For a mini-batch $\mathcal{B}$ with image and text feature matrices $X_{\mathcal{B}}$ and $Y_{\mathcal{B}}$, their continuous outputs are

$$P_I = f_I(\mathcal{N}(X_{\mathcal{B}})), \quad P_T = f_T(\mathcal{N}(Y_{\mathcal{B}})). \tag{11}$$

Training targets are formed from the fused and modality-local continuous representations produced by GJHL. Continuous and binarized versions of these targets are balanced according to training progress, allowing the modality-specific hash functions to remain guided by the continuous teacher while progressively approaching binary-compatible outputs. The detailed target construction is provided in the Appendix B.4.

The modality-specific hash functions are optimized by

$$\mathcal{L}_{\mathrm{hash}} = \mathcal{L}_{\mathrm{DISTILL}} + \lambda_r\mathcal{L}_{\mathrm{RANK}}, \tag{12}$$

where $\mathcal{L}_{\mathrm{DISTILL}}$ is a composite teacher-guided objective that aligns the modality-specific outputs with the GJHL-derived training targets while jointly maintaining paired cross-modal consistency and preserving the similarity relations specified by $S_{\mathcal{B}}^e$. The separately constructed graph-ranking objective $\mathcal{L}_{\mathrm{RANK}}$ further encourages cross-modal rankings consistent with the current similarity matrix, with $\lambda_r \geq 0$ controlling its contribution. The internal components and weighting of $\mathcal{L}_{\mathrm{DISTILL}}$ are detailed in the Appendix B.4.

Within each epoch, GJHL is optimized under the current similarity matrix S^e before the modality-specific hash functions are updated. Only after this update can their continuous outputs contribute to a scheduled similarity-matrix update, and any newly estimated similarity matrix is used only in subsequent epochs.

CSCI, HASR, and GJHL are used only during training and are not involved in inference. At inference time, only the modality-specific hash functions f_I and f_T are used for retrieval.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

❶ **Datasets.** We evaluated CSCIH on three widely used multi-label image-text benchmark datasets: MIRFLICKR-25K (Huiskes & Lew, 2008), NUS-WIDE (Chua et al., 2009), and MS-COCO (Lin et al., 2014). MIRFLICKR-25K contains 24,581 image-text pairs annotated with 24 classes, while NUS-WIDE contains 195,834 pairs covering 21 classes. For MS-COCO, we merged the training and validation sets, resulting in 122,218 paired samples from 80 classes. We used 10,000, 10,500, and 10,000 pairs for training on MIRFLICKR-25K, NUS-WIDE, and MS-COCO, respectively. The corresponding query sets contain 2,000, 2,100, and 5,000 pairs, respectively, and the retrieval databases contain 22,581, 193,734, and 117,218 pairs.

❷ **Baselines.** To evaluate the effectiveness of our proposed CSCIH method, we compared it with eight representative unsupervised cross-modal hashing methods, including UCCH (Hu et al., 2023), CIRH (Zhu et al., 2023), CDTH (Li et al., 2024a), DUHEG (Song et al., 2025), VTMUCH (Fan & Cao, 2025), MSSH (Chen et al., 2026c), CTHRS (Yang et al., 2026b), and EGICH (Chen et al., 2026b).

❸ **Evaluation Metrics.** We adopted the widely used mean average precision (mAP) as the primary evaluation metric and reported the mAP@all results for all methods. In addition, precision-recall (PR) curves were used to evaluate ranking performance at different recall levels. All methods were evaluated on image-to-text (I2T) and text-to-image (T2I) retrieval with hash code lengths of 16, 32, and 64 bits.

❹ **Implementation Details.** All methods adopt image and text feature extraction through CLIP ViT-B/32 (Radford et al., 2021). All experiments were conducted on an NVIDIA GeForce RTX 4090 GPU with the random seed fixed to 2026. For our proposed CSCIH method, all components were optimized using the Adam optimizer. Specifically, we use learning rates of 5×10^{-4} for CSCI, 1×10^{-3} for the main network, and 1×10^{-4} for both image and text retrieval heads. The model is trained for 40 epochs with a batch size of 256 and the CSCI uses 8 environment prototypes. For MIRFlickr-25K, NUS-WIDE, and MS-COCO, the graph neighborhood ratios are set to 0.08, 0.04, and 0.08, the graph refresh blending coefficients to 0.85, 0.95, and 0.85, and the GJHL ranking weights to 0.70, 0.65, and 0.55, respectively. The NBR and quantization loss weights are fixed to 10.0 and 0.55 across all datasets. At inference time, retrieval is performed using only the input from the query modality, without requiring paired input or information from the other modality.

5.2 COMPARISON WITH THE BASELINES

Table 1 reports the mAP results of CSCIH and the competing methods on MIRFlickr-25K, NUS-WIDE, and MS-COCO. CSCIH consistently outperforms the competing methods across all three datasets, retrieval directions, and hash-code lengths, demonstrating its effectiveness in learning discriminative cross-modal hash representations under evolving structural supervision. The consistent gains across different retrieval settings indicate that the improvement is not limited to a particular modality direction or code capacity, but remains stable under varying retrieval conditions. Furthermore, Fig. 2 presents the PR and Top K Precision curves with 32-bit hash codes. CSCIH maintains favorable performance across different recall levels and retrieval depths, showing that its advantage extends beyond aggregated mAP to overall ranking quality. Additional baseline comparisons and complementary analyses are provided in Appendix C.1.

Table 1: Comparison of mAP performance (%) across various cross-modal hashing methods.

Task	Method	Reference	MIRFlickr-25K			NUS-WIDE			MS-COCO		
			16	32	64	16	32	64	16	32	64
I→T	UGACH	AAAI 2018	67.0	68.6	69.1	54.1	57.0	57.0	54.2	56.4	58.3
	DJSRH	ICCV 2019	60.7	63.1	63.8	39.5	43.7	45.5	43.9	50.0	54.6
	UDIH	IJCNN 2020	68.2	68.8	70.1	54.8	56.2	59.8	55.5	57.1	59.4
	JDSH	SIGIR 2020	64.9	66.1	66.5	54.6	56.0	57.5	49.8	53.0	55.4
	DGCPN	AAAI 2021	70.3	72.0	72.6	60.3	63.3	63.8	55.4	59.4	58.1
	DAEH	TCSVT 2022	69.0	71.1	72.6	55.7	60.9	61.0	52.7	53.6	54.4
	CIRH	TKDE 2023	68.9	69.8	69.8	60.1	61.2	61.9	58.9	62.8	64.3
	DUHEG	ICML 2025	70.2	72.3	71.9	44.9	45.3	46.0	54.1	58.6	60.5
	MSSH	Inf. Fusion 2026	69.7	70.1	71.1	53.8	58.0	59.1	50.2	50.4	51.6
	CTHRS	TGRS 2026	65.1	66.4	68.7	49.7	51.7	53.7	59.4	59.6	61.3
	CSCIH	Ours	**71.2**	**74.1**	**74.8**	**62.8**	**65.9**	**67.2**	**59.5**	**64.5**	**64.8**
T→I	UGACH	AAAI 2018	67.6	68.7	69.1	53.5	56.0	56.5	54.0	56.2	58.0
	DJSRH	ICCV 2019	61.0	62.4	63.7	40.0	43.9	46.3	43.9	50.2	54.5
	UDIH	IJCNN 2020	67.8	69.4	70.3	55.3	57.1	60.7	55.2	56.0	58.5
	JDSH	SIGIR 2020	65.2	64.9	67.1	54.2	58.8	55.7	48.8	53.7	53.6
	DGCPN	AAAI 2021	70.3	72.3	73.6	60.8	64.0	64.2	55.3	59.1	57.6
	DAEH	TCSVT 2022	71.1	72.7	74.7	56.8	62.4	62.6	53.4	54.1	55.0
	CIRH	TKDE 2023	70.2	71.1	70.7	61.2	62.1	62.9	59.2	63.5	64.9
	DUHEG	ICML 2025	69.3	71.9	72.3	49.2	52.2	51.6	53.3	58.8	60.1
	MSSH	Inf. Fusion 2026	70.3	70.7	71.0	54.5	58.2	59.1	49.7	49.7	51.7
	CTHRS	TGRS 2026	64.8	66.4	68.6	50.2	52.1	54.4	59.4	59.7	61.3
	CSCIH	Ours	**72.1**	**75.2**	**75.6**	**63.0**	**66.7**	**67.6**	**59.7**	**64.7**	**65.2**

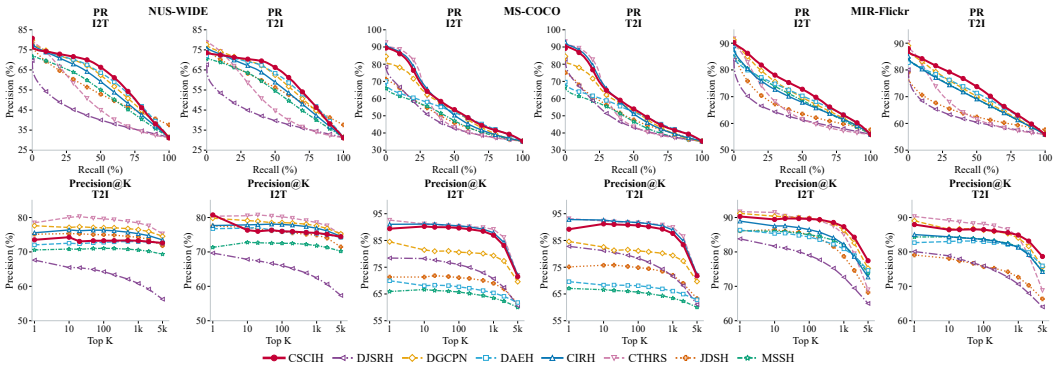


Figure 2: Retrieval performance comparison on MIRFLICKR-25K, NUS-WIDE, and MS-COCO under image-to-text and text-to-image settings, measured by precision-recall curves and Precision@K.

5.3 Ablation Study

To evaluate the contribution of the main components in CSCIH, we conduct additive ablation experiments on MIRFlickr-25K, NUS-WIDE, and MS-COCO. As shown in Table 2, CSCI, HASR, and GJHL are progressively introduced into the baseline model. CSCIH-V1 incorporates CSCI for subject-context decomposition and intervention-based refinement, CSCIH-V2 further introduces HASR to regulate the re-entry of learned structural relations, and CSCIH-V3 additionally includes GJHL to provide complementary geometric organization for hash learning. CSCI yields clear improvements on MIRFlickr-25K and NUS-WIDE, while its isolated effect is more limited on MS-COCO. Adding HASR further improves retrieval performance across the three datasets, and the subsequent inclusion of GJHL produces consistent gains in all settings, with particularly evident improvements on MS-COCO. Overall, the full CSCIH achieves the strongest performance across both

Table 2: Ablation study of CSCIH (%) on MIRFLICKR-25K, NUS-WIDE, and MS-COCO.

Methods	CSCI	HASR	GJHL	MIRFLICKR-25K		NUS-WIDE		MS-COCO	
				I→T	T→I	I→T	T→I	I→T	T→I
Baseline	×	×	×	68.9	70.2	61.2	62.1	62.8	63.5
CSCIH-V1	√	×	×	71.9	72.9	64.4	64.6	62.2	62.3
CSCIH-V2	√	√	×	73.5	74.5	64.7	65.0	62.9	62.7
CSCIH-V3	√	√	√	**74.1**	**74.9**	**65.3**	**66.2**	**64.5**	**64.7**

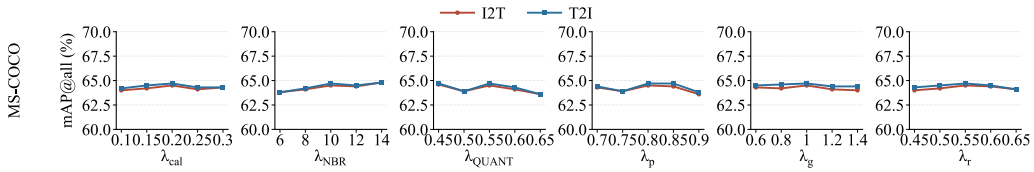


Figure 3: Sensitivity analysis of key hyperparameters of CSCIH on MS-COCO with 32-bit hash codes.

retrieval directions, supporting the complementary roles of the three components in the complete framework. More detailed ablation studies and analyses are provided in Appendix C.2.

5.4 Parameter Analysis

We evaluate the sensitivity of six key hyperparameters on MS-COCO with 32-bit hash codes. Specifically, λ_{cal} controls the CSCI calibration regularization, λ_{NBR} and λ_{QUANT} weight neighborhood reconstruction and quantization in continuous teacher learning, while λ_p, λ_g, and λ_r control paired cross-modal consistency, similarity-relation preservation, and graph-ranking supervision in modality-specific hash learning, respectively. As shown in Fig. 3, retrieval performance varies only modestly across the tested ranges, with the two retrieval directions exhibiting closely aligned trends. Although several parameters show shallow local optima, no substantial degradation is observed within the evaluated ranges. Overall, CSCIH remains stable under different hyperparameter settings, indicating limited sensitivity to their specific choices. Additional sensitivity analyses on MIRFlickr-25K and NUS-WIDE are provided in Appendix C.3, where similar trends are observed across the two datasets.

6 Conclusion

In this paper, we propose an unsupervised cross-modal hashing method, termed CSCIH, to address similarity estimation under subject-context entanglement. It is designed to reduce context-induced associations that may be incorrectly encoded as semantic similarities and further propagated through successive similarity updates. Specifically, the CSCI module learns subject- and context-oriented representations, performs an approximate differentiable intervention, and evaluates subject consistency across alternative contextual states. Based on this consistency, HASR adaptively balances similarity relations derived from subject-oriented and continuous hash representations over the course of training, while retaining the initial similarity matrix as an anchor to alleviate accumulated drift. In addition, GJHL preserves modality-local relational structures and models cross-modal relations in complementary Euclidean and hyperbolic spaces, integrating them into a continuous teacher representation for modality-specific hash learning. Experiments on three benchmark datasets show consistent improvements over representative unsupervised cross-modal hashing methods across retrieval directions and hash-code lengths, while also validating the effectiveness of the proposed components. These results highlight the importance of considering how similarity relations are formed and progressively incorporated into supervision in unsupervised cross-modal hashing.

7 Ethics Statement

Our research adheres to the ICLR Code of Ethics. All datasets used in this study are publicly available. The code and related materials will be appropriately released to ensure transparency and reproducibility of our work.

References

Dong Chen, Miaomiao Cheng, Chen Min, and Liping Jing. Unsupervised deep imputed hashing for partial cross-modal retrieval. In *IJCNN*, 2020.

Haipeng Chen, Yu Liu, Xun Yang, Yuheng Liang, and Yingda Lyu. Causality-aligned semantic recovery for incomplete cross-modal retrieval. In *AAAI*, 2026a.

Jiali Chen, Ruitao Pu, Dezhong Peng, Xiaomin Song, Yingke Chen, and Yuan Sun. External guidance incomplete cross-modal hashing. *IEEE Transactions on Image Processing*, 2026b.

Yunfei Chen, Yitian Long, Zhan Yang, and Jun Long. Multi-dimensional semantic self-learning hashing for multimedia retrieval. *Information Fusion*, 2026c.

Zhiwei Chen, Yupeng Hu, Zhiheng Fu, Zixu Li, Jiale Huang, Qinlei Huang, and Yinwei Wei. INTENT: Invariance and discrimination-aware noise mitigation for robust composed image retrieval. In *AAAI*, 2026d.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *CIVR*, 2009.

Haozhi Fan and Yuan Cao. Vision-guided text mining for unsupervised cross-modal hashing with community similarity quantization. In *AAAI*, 2025.

Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *MIR*, 2008.

Mingjin Kuai, Jun Long, and Zhan Yang. Statistical model-driven similarity hashing: Bridging modalities for efficient unsupervised retrieval. In *AAAI*, 2025.

Fengling Li, Bowen Wang, Lei Zhu, Jingjing Li, Zheng Zhang, and Xiaojun Chang. Cross-domain transfer hashing for efficient cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024a.

Xiutian Li, Siqi Sun, and Rui Feng. Causal representation learning via counterfactual intervention. In *AAAI*, 2024b.

Yanmin Li, Zhilong Mao, Mao Wang, Lihua Liu, Jibing Wu, and Weidong Bao. Prototype-based causal intervention for multi-label image classification. In *CVPR*, 2026.

Xusheng Liang, Lihua Zhou, Nianxin Li, Miao Xu, Ziyang Song, Dong Yi, Jinlin Wu, Jiawei Ma, Hongbin Liu, Zhen Lei, and Jiebo Luo. Multimodal causality-driven representation learning for generalizable medical image segmentation. In *CVPR*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Ruyang Liu, Hao Liu, Ge Li, Haodi Hou, TingHao Yu, and Tao Yang. Contextual debiasing for visual recognition with causal mechanisms. In *CVPR*, 2022.

Song Liu, Shengsheng Qian, Yang Guan, Jiawei Zhan, and Long Ying. Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval. In *SIGIR*, 2020.

Saiyang Na, Feng Jiang, Qifeng Zhou, Wenliang Zhong, Thao M. Dang, Yuzhi Guo, Hehuan Ma, Chunyuan Li, Weizhi An, and Junzhou Huang. HyperGRAM: Hyperbolic gramian volumes for multimodal alignment. In *CVPR*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *CVPR*, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

Qihong Song, Xiting Liu, Hongyuan Zhu, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Deep unsupervised hashing via external guidance. In *ICML*, 2025.

Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *ICCV*, 2019.

Jingyao Wang, Siyu Zhao, Wenwen Qiang, Jiangmeng Li, Changwen Zheng, Fuchun Sun, and Hui Xiong. Towards the causal complete cause of multi-modal representation learning. In *ICML*, 2025.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: A systematic review of methods and future directions. *Proceedings of the IEEE*, 2024a.

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Xue Wang, Zheng Guan, Wenhua Qian, Chengchao Wang, and Runzhuo Ma. Multi-modal image fusion via intervention-stable feature learning. In *CVPR*, 2026.

Shaowu Xu, Xibin Jia, Chao Fan, Junyu Gao, Jing Chang, and Qianmei Sun. Progressive cross-modal causal intervention for long-term action recognition. In *CVPR*, 2026.

Dejie Yang, Dayan Wu, Wanqian Zhang, Haisu Zhang, Bo Li, and Weiping Wang. Deep semantic-alignment hashing for unsupervised cross-modal retrieval. In *ICMR*, 2020.

Fan Yang, Yuanzhi Zhao, Haimei Zhao, Yudong Zhao, and Haikun Xu. Mask to align, weight to disambiguate: Reliable unsupervised cross-modal hashing with masked-weight contrast. In *CVPR*, 2026a.

Menglin Yang, Min Zhou, Rex Ying, Yankai Chen, and Irwin King. Hyperbolic representation learning: Revisiting and advancing. In *ICML*, 2023.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *CVPR*, 2021a.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *CVPR*, 2021b.

Zhan Yang, Weilu Long, Binghong Chen, Longzhi Sun, Yinan Li, and Xiaoyao Sun. Unsupervised contrastive transformer hashing for cross-modal remote sensing retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2026b.

Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *AAAI*, 2021.

Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. In *AAAI*, 2018.

Lei Zhu, Xize Wu, Jingjing Li, Zheng Zhang, Weili Guan, and Heng Tao Shen. Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

A ALGORITHM

Algorithm 1 summarizes the epoch-wise training procedure of CSCIH. Within each epoch, CSCI produces refined modality representations and subject consistency statistics, and GJHL constructs the continuous teacher representations under the current shared and modality-local similarity matrices. The cached teacher representations then guide the modality-specific hash functions. At scheduled updates, HASR refreshes the shared and modality-local similarity matrices using the current retrieval representations, subject oriented representations, and subject consistency score. Only f_I and f_T are retained for inference.

Algorithm 1 Training Procedure of CSCIH

Require: Paired dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$; hash length c
Ensure: Hash functions $f_I(\cdot)$ and $f_T(\cdot)$

Initialize S^0, S_I^0, S_T^0
$S_I \leftarrow S_I^0$, $S_T \leftarrow S_T^0$
for each epoch e **do**
  for each mini-batch $\mathcal{B}$ **do**
    $(\bar{Z}_{\mathcal{B}}^I, \bar{Z}_{\mathcal{B}}^T, h_{\mathcal{B}}) \leftarrow \mathrm{CSCI}(X_{\mathcal{B}}, Y_{\mathcal{B}})$
    $(H_I, H_J, H_T, H) \leftarrow \mathrm{GJHL}(\bar{Z}_{\mathcal{B}}^I, \bar{Z}_{\mathcal{B}}^T, S_{\mathcal{B}}^e, S_I, S_T)$
    Update CSCI and GJHL with $\mathcal{L}_{\mathrm{teacher}}$
    Cache $\mathrm{sg}(H_I)$, $\mathrm{sg}(H_T)$, $\mathrm{sg}(H)$
  end for
  $h_e \leftarrow |\mathcal{B}_e|^{-1} \sum_{\mathcal{B} \in \mathcal{B}_e} h_{\mathcal{B}}$
  for each cached mini-batch $\mathcal{B}$ **do**
    $P_I \leftarrow f_I(\mathcal{N}(X_{\mathcal{B}}))$, $P_T \leftarrow f_T(\mathcal{N}(Y_{\mathcal{B}}))$
    Update f_I, f_T with $\mathcal{L}_{\mathrm{hash}}$
  end for
  if a similarity-matrix update is scheduled after epoch e **then**
    Obtain $P_I^e, P_T^e, U_s^I, U_s^T$ over $\mathcal{D}$
    $(S^{e+1}, S_I, S_T) \leftarrow \mathrm{HASR}(S^0, S_I^0, S_T^0, P_I^e, P_T^e, U_s^I, U_s^T, h_e)$
  else
    $(S^{e+1}, S_I, S_T) \leftarrow (S^e, S_I, S_T)$
  end if
end for
return $f_I(\cdot), f_T(\cdot)$

B DETAILED METHOD DERIVATIONS

This section provides the detailed formulations of CSCI, HASR, GJHL, and teacher-guided modality-specific hash learning, using the notation introduced in Sections 4.1-4.4. Unless otherwise stated, c denotes the hash-code length, $\kappa > 0$ the hyperbolic curvature magnitude, $\mathrm{Sym}(A) = (A + A^\top)/2$ matrix symmetrization, and $[\cdot]_{[a,b]}$ element-wise clipping to $[a, b]$.

B.1 CAUSAL SUBJECT-CONTEXT INTERVENTION

For modality $m \in \{I, T\}$, let $z_i^I = x_i$ and $z_i^T = y_i$. CSCI learns a sample-dependent gate $g_i^m = \sigma(G_m(z_i^m))$ and derives subject and context oriented representations as

$$s_i^m = E_s^m(g_i^m \odot z_i^m), \quad c_i^m = E_c^m((1 - g_i^m) \odot z_i^m).$$

The paired image and text representations are further fused through

$$\Phi_r(a, b) = F_r\big([a; b; \tfrac{1}{2}(a + b); |a - b|]\big), \quad r \in \{s, c\},$$

yielding

$$s_i = \Phi_s(s_i^I, s_i^T), \quad c_i = \Phi_c(c_i^I, c_i^T).$$

The modality-specific subject oriented representations s_i^I and s_i^T are retained for constructing similarity relations, whereas the shared representations s_i and c_i are used to assess subject consistency under contextual variation.

CSCI performs the approximate differentiable intervention by restricting residual refinement to context-oriented dimensions while preserving subject oriented ones:

$$\bar{z}_i^m = z_i^m + \eta_e(1 - g_i^m) \odot \tanh(D_m(c_i)),$$

where $\eta_e \geq 0$ controls the refinement strength and D_m maps the shared context-oriented representation back to the representation space of modality m. Since the residual is modulated by $(1 - g_i^m)$ and $|\tanh(\cdot)| \leq 1$, larger subject oriented gate values produce smaller perturbations, thereby concentrating residual refinement on context-oriented dimensions.

To characterize contextual variation, CSCI maps the shared context-oriented representation to a distribution over K latent contextual-state prototypes $\{p_k\}_{k=1}^K$:

$$\pi_{ik} = \frac{\exp\big(\mathcal{N}(c_i)^\top \mathcal{N}(p_k)/\tau_{\text{env}}\big)}{\sum_{\ell=1}^K \exp(\mathcal{N}(c_i)^\top \mathcal{N}(p_\ell)/\tau_{\text{env}})},$$

where $\pi_i = (\pi_{i1}, \ldots, \pi_{iK})$ denotes the contextual-state distribution of sample i. The learned decomposition is stabilized by a compact calibration regularizer that promotes cross-modal subject alignment, reduces redundancy between subject and context oriented representations, and encourages confident yet non-collapsed contextual-state assignments. For a mini-batch B, we define

$$\ell_{\text{align}} = \frac{1}{|B|} \sum_{i \in B} \big(1 - \cos(s_i^I, s_i^T)\big)^2,$$

and

$$\ell_{\text{sep}} = \frac{1}{2|B|} \sum_{i \in B} \left[\cos^2(s_i^I, c_i^I) + \cos^2(s_i^T, c_i^T)\right].$$

The former aligns paired subject oriented representations across modalities, while the latter discourages redundancy between subject and context oriented representations within each modality. Let

$$\bar{\pi}_B = \frac{1}{|B|} \sum_{i \in B} \pi_i, \quad u = \left(\tfrac{1}{K}, \ldots, \tfrac{1}{K}\right).$$

The contextual-state penalties are

$$\ell_{\text{conf}} = \frac{1}{|B| \log K} \sum_{i \in B} H(\pi_i), \quad \ell_{\text{bal}} = \|\bar{\pi}_B - u\|_2^2.$$

Minimizing ℓ_{conf} encourages confident sample-level assignments, while ℓ_{bal} discourages collapse of the mini-batch distribution to a small subset of contextual states. The calibration regularizer is therefore

$$\mathcal{L}_{\text{cal}} = \frac{\lambda_{\text{cal}}(e)}{4} \left(\ell_{\text{align}} + \ell_{\text{sep}} + \ell_{\text{conf}} + \ell_{\text{bal}}\right),$$

where $\lambda_{\text{cal}}(e) \geq 0$ denotes the effective epoch-dependent calibration weight.

CSCI further assesses whether subject information remains consistent across alternative contextual states. For each sample i, candidate samples are drawn from the current mini-batch and a detached pool accumulated from preceding mini-batches within the current epoch. Subject similarity and contextual-state discrepancy are measured by

$$a_{ij} = \mathcal{N}(s_i)^\top \mathcal{N}(s_j), \quad d_{ij}^{\text{env}} = \frac{\text{JS}(\pi_i, \pi_j)}{\log 2},$$

respectively. Let τ_s^e and τ_d^e denote the effective epoch-dependent thresholds at epoch e, determined by the current CSCI training progress. The eligible contextual alternatives for sample i are defined as

$$\mathcal{A}_i^e = \left\{j \neq i : a_{ij} \geq \tau_s^e,\ d_{ij}^{\text{env}} \geq \tau_d^e\right\}.$$

A candidate therefore qualifies only when it preserves sufficiently similar subject information while exhibiting a sufficiently different contextual state. These alternatives are used only for subject consistency assessment and are not injected into the refined representation. The residual refinement of sample i is computed solely from its own context-oriented representation c_i.

Let

$$r_B = \frac{|\{i \in B : \mathcal{A}_i^e \neq \varnothing\}|}{|B|},$$

which denotes the proportion of samples in mini-batch B for which at least one eligible contextual alternative can be identified. CSCI combines this availability with the confidence of contextual-state assignments:

$$h_B = \lambda_h r_B + (1 - \lambda_h) \left(1 - \frac{1}{|B| \log K} \sum_{i \in B} H(\pi_i)\right).$$

Averaging over the mini-batches of epoch e yields

$$h_e = \frac{1}{|\mathcal{B}_e|} \sum_{B \in \mathcal{B}_e} h_B.$$

Here, h_e is the epoch-wise subject consistency score. The modality-specific subject oriented representations are subsequently used to construct similarity relations for HASR, while h_e regulates the contribution of these relations at the subsequent similarity-matrix update.

B.2 Health-Aware Structure Refresh

For paired representation matrices $U, V \in \mathbb{R}^{N \times d}$, HASR constructs

$$R_{\text{intra}} = a_1 \mathcal{N}(U)\mathcal{N}(U)^\top + (1 - a_1)\mathcal{N}(V)\mathcal{N}(V)^\top, \quad R_{\text{cm}} = \frac{1}{2}\left[\mathcal{N}(U)\mathcal{N}(V)^\top + \mathcal{N}(V)\mathcal{N}(U)^\top\right],$$

and

$$R = (1 - \gamma_{\text{cm}}) R_{\text{intra}} + \gamma_{\text{cm}} R_{\text{cm}}.$$

Here, a_1 balances the two intra-modal components and $\gamma_{\text{cm}} \in [0, 1]$ controls the direct cross-modal component. Let

$$R_k = \text{Sym}(\mathcal{P}_\rho(R)),$$

where $\mathcal{P}_\rho$ retains the strongest row-wise neighbors according to graph density ρ. The shared relation constructor is

$$\mathcal{K}(U, V) = \text{Diag}_1\left([a_2 R_k + (1 - a_2)(2\sigma(R_k) - 1)]_{[-1,1]}\right),$$

where a_2 balances the retained and sigmoid-smoothed similarities, and $\text{Diag}_1(\cdot)$ restores self-similarity.

The fixed initial similarity matrix S^0 is constructed once from the initial training representations using the same relation constructor. At a scheduled update following epoch e,

$$S^e_{\text{hash}} = \mathcal{K}(P_I^e, P_T^e), \quad S^e_{\text{sub}} = \mathcal{K}(U_s^I, U_s^T).$$

HASR balances these complementary similarity relations according to subject consistency and training progress:

$$\omega_e = [\lambda_s \alpha_e h_e]_{[0,1]}, \quad \widetilde{S}^e = (1 - \omega_e) S^e_{\text{hash}} + \omega_e S^e_{\text{sub}},$$

where $\lambda_s \geq 0$ is the subject-relation scale. To constrain accumulated drift, the next similarity matrix is anchored to S^0:

$$S^{e+1} = \mathcal{T}\left(\lambda_0 S^0 + (1 - \lambda_0)\widetilde{S}^e\right),$$

where $\lambda_0 \in [0, 1]$ and

$$\mathcal{T}(A) = \text{Diag}_1\left(\left[\text{Sym}\left(\mathcal{P}_\rho^u(\text{Sym}(A))\right)\right]_{[-1,1]}\right).$$

Here, $\mathcal{P}_\rho^u$ optionally reapplies row-wise sparsification during the update and is the identity when this operation is disabled.

CSCIH also maintains modality-local similarity matrices S_I and S_T for GJHL. Let $\mathcal{K}_m(U)$ apply the same sparsification, smoothing, symmetrization, and diagonal restoration to the single-modal similarity $\mathcal{N}(U)\mathcal{N}(U)^\top$. With S_m^0 denoting the corresponding fixed initial matrix,

$$\widetilde{S}_m^e = (1 - \omega_e)\mathcal{K}_m(P_m^e) + \omega_e \mathcal{K}_m(U_s^m),$$

and

$$S_m \leftarrow \mathcal{T}\left(\lambda_0 S_m^0 + (1-\lambda_0)\widetilde{S}_m^e\right), \quad m \in \{I, T\}.$$

For graph propagation, a similarity matrix S is converted to

$$\widetilde{A} = \varphi_g\left(\mathrm{Sym}(S) - \mathrm{Diag}(S)\right) + I, \quad \mathcal{G}(S) = D^{-1/2}\widetilde{A}D^{-1/2},$$

where

$$D_{ii} = \max\left(\sum_j \widetilde{A}_{ij}, \epsilon\right).$$

Thus,

$$A_{\mathcal{B}}^e = \mathcal{G}(S_{\mathcal{B}}^e), \quad A_m = \mathcal{G}(S_m), \quad m \in \{I, T\},$$

where $S_{\mathcal{B}}^e$ is the mini-batch submatrix of the current shared similarity matrix. The former supports cross-modal relational modeling, while the latter preserves modality-local relational structures in GJHL.

If no similarity-matrix update is scheduled after epoch e, the current matrices remain unchanged. Otherwise, the updated matrices are constructed after completing epoch e and are used only in subsequent optimization epochs.

B.3 Geometry-Aware Joint Hash Learning

For modality $m \in \{I, T\}$, let $n_1(i)$ and $n_2(i)$ denote the two strongest local neighbors of sample i selected from S_m, with $n_0(i) = i$. The adaptive neighbor gate receives

$$q_i^m = \left[\bar{z}_i^m;\, \bar{z}_{n_1(i)}^m;\, \bar{z}_{n_2(i)}^m;\, \bar{z}_{n_1(i)}^m - \bar{z}_i^m;\, \bar{z}_{n_2(i)}^m - \bar{z}_i^m;\, \cos(\bar{z}_i^m, \bar{z}_{n_1(i)}^m);\, \cos(\bar{z}_i^m, \bar{z}_{n_2(i)}^m)\right]$$

and predicts

$$\widehat{\mathbf{w}}_i^m = \mathrm{softmax}(G_N^m(q_i^m)).$$

The weights are introduced progressively according to training progress and satisfy $w_{ij}^m \geq 0$ and $\sum_{j=0}^2 w_{ij}^m = 1$. The neighbor-mixed representation is

$$\tilde{z}_i^m = \sum_{j=0}^{2} w_{ij}^m \bar{z}_{n_j(i)}^m.$$

Let

$$E_m = \mathcal{N}\Big(\mathrm{Enc}_m(\tilde{Z}_m)\Big),$$

where $\tilde{Z}_m$ stacks the neighbor-mixed representations. Using $A_m = \mathcal{G}(S_m)$ from Appendix B.2, the modality-local representation is

$$V_m = \phi_m(A_m E_m W_m), \quad m \in \{I, T\}.$$

For cross-modal relational modeling, define

$$C_{\mathrm{raw}} = E_I E_T^\top - \mathrm{Diag}(E_I E_T^\top), \quad C_{\mathrm{sim}} = \mathcal{P}_{k_c}^{\leftrightarrow}(C_{\mathrm{raw}}),$$

where $\mathcal{P}_{k_c}^{\leftrightarrow}$ retains the union of the strongest image-to-text and text-to-image links. These links are combined with paired identity correspondence:

$$C_e = \lambda_{\mathrm{id}}(e) I + (1 - \lambda_{\mathrm{id}}(e))\, C_{\mathrm{sim}}.$$

The joint graph is

$$A_{\mathrm{cm}}^e = \begin{bmatrix} A_{\mathcal{B}}^e & C_e \\ C_e^\top & A_{\mathcal{B}}^e \end{bmatrix}, \quad A_{\mathcal{B}}^e = \mathcal{G}(S_{\mathcal{B}}^e).$$

The Euclidean branch provides the base joint representation. With $E_C = [E_I; E_T]$,

$$\begin{bmatrix} U_I^E \\ U_T^E \end{bmatrix} = A_{\mathrm{cm}}^e E_C W_J, \quad V_J^E = \phi_J(U_I^E + U_T^E).$$

The complementary hyperbolic branch operates in the Poincaré ball

$$\mathbb{D}_\kappa^d = \left\{ u \in \mathbb{R}^d : \kappa \|u\|_2^2 < 1 \right\},$$

where $\kappa > 0$ is the curvature magnitude. For the stacked branch input E_C^H,

$$U_0 = \exp_0^\kappa \left(\Psi_{\text{in}}(E_C^H) \right), \quad T_0 = \log_0^\kappa(U_0),$$

and

$$T_1 = \text{BN}_{\text{hyp}} \left(A_{\text{cm}}^e T_0 W_H \right), \quad U_1 = \exp_0^\kappa \left(\Psi_{\text{prop}}(T_1) \right).$$

After splitting U_1 into U_1^I and U_1^T, the hyperbolic joint representation is

$$V_J^H = \log_0^\kappa \left(\mathcal{R}_\kappa(U_1^I) \oplus_\kappa \mathcal{R}_\kappa(U_1^T) \right).$$

The final joint representation is

$$V_J = V_J^E + \lambda_H(e) \left(V_J^H - V_J^E \right), \quad \lambda_H(e) \in [0, 1],$$

where $\lambda_H(e)$ controls the effective epoch-dependent contribution of the complementary hyperbolic branch.

The modality-local representations V_I, V_T and the joint representation V_J are mapped to

$$H_I = \tanh\big(\text{BN}_I \left(W_I^H [V_I; V_J] \right)\big),$$
$$H_J = \tanh\big(\text{BN}_J \left(W_J^H V_J \right)\big),$$
$$H_T = \tanh\big(\text{BN}_T \left(W_T^H [V_J; V_T] \right)\big).$$

For bit-wise fusion, let

$$q_H = [H_I;\, H_J;\, H_T;\, H_I - H_J;\, H_T - H_J;\, |H_I - H_T|;\, c_{IJ};\, c_{JT};\, c_{IT}],$$

where $c_{IJ} = \cos(H_I, H_J)$, $c_{JT} = \cos(H_J, H_T)$, and $c_{IT} = \cos(H_I, H_T)$. The branch weights and primary fused representation are

$$\Gamma = \text{softmax}_{\text{branch}} \left(G_H(q_H) + \Pi_e \right), \quad \Gamma_I + \Gamma_J + \Gamma_T = \mathbf{1},$$

and

$$H_F = \Gamma_I \odot H_I + \Gamma_J \odot H_J + \Gamma_T \odot H_T.$$

We further define

$$M_H = \frac{1}{2}(H_I + H_T), \quad R_H = \text{BN}_R \left(W_R \left[H_F;\, M_H;\, H_I \odot H_T;\, |H_I - H_T| \right] \right),$$

yielding the continuous teacher representation

$$H = \tanh \left(\text{BN}_H \left(W_C [H_F; M_H; R_H] + H_F \right) \right).$$

For neighborhood reconstruction,

$$\widehat{X}_I = D_I(H + H_I), \quad \widehat{X}_T = D_T(H + H_T).$$

Let $X_I = X_{\mathcal{B}}$ and $X_T = Y_{\mathcal{B}}$. The objective is

$$\mathcal{L}_{\text{NBR}} = \sum_{m \in \{I,T\}} \mathcal{E}(\widehat{X}_m, X_m) + \lambda_n \sum_{m \in \{I,T\}} \mathcal{E}(\widehat{X}_m, \tilde{Z}_m).$$

For binary compatibility, define

$$Q(H) = \begin{cases} \text{sign}(H), & \text{hard mode,} \\ \tanh(H/\tau_q), & \text{soft mode,} \end{cases}$$

with

$$\mathcal{L}_{\text{QUANT}} = \mathcal{E} \left(H, \text{sg}(Q(H)) \right).$$

The teacher-side objective is therefore

$$\mathcal{L}_{\text{teacher}} = \lambda_{\text{NBR}} \mathcal{L}_{\text{NBR}} + \mathcal{L}_{\text{cal}} + \lambda_{\text{QUANT}} \mathcal{L}_{\text{QUANT}}.$$

The resulting H serves as the continuous teacher representation for the modality-specific hash learning described next.

B.4 Teacher-Guided Modality-Specific Hash Learning

Given the fused representation H and modality-local representations H_I, H_T from GJHL, we construct

$$T_{\mathrm{sh}}^{\mathrm{soft}} = \left[(1-\beta_{\mathrm{cross}})H + \beta_{\mathrm{cross}} \tanh\left(\frac{H_I + H_T}{2} \right) \right]_{[-1,1]},$$

and

$$T_m^{\mathrm{soft}} = \left[(1-\beta_{\mathrm{mod}})T_{\mathrm{sh}}^{\mathrm{soft}} + \beta_{\mathrm{mod}} H_m \right]_{[-1,1]}, \quad m \in \{I, T\},$$

where $\beta_{\mathrm{cross}}, \beta_{\mathrm{mod}} \in [0,1]$ control the cross-modal and modality-specific contributions.

Continuous and binary-compatible targets are balanced according to training progress:

$$T_m^e = \mathrm{sg}\left(\left[(1-\zeta_e)T_m^{\mathrm{soft}} + \zeta_e \operatorname{sign}(T_m^{\mathrm{soft}}) \right]_{[-1,1]} \right),$$

and

$$T_H^e = \mathrm{sg}\left(\left[(1-\zeta_e)H + \zeta_e \operatorname{sign}(H) \right]_{[-1,1]} \right),$$

where ζ_e increases with training progress.

For mini-batch $\mathcal{B}$,

$$P_I = f_I(\mathcal{N}(X_{\mathcal{B}})), \quad P_T = f_T(\mathcal{N}(Y_{\mathcal{B}})).$$

Teacher fitting is

$$\mathcal{L}_{\mathrm{fit}} = \mathcal{E}(P_I, T_I^e) + \mathcal{E}(P_T, T_T^e) + \lambda_{\mathrm{sh}} \left[\mathcal{E}(P_I, T_H^e) + \mathcal{E}(P_T, T_H^e) \right].$$

The composite distillation objective further preserves paired consistency and the current similarity relations:

$$\mathcal{L}_{\mathrm{DISTILL}} = \mathcal{L}_{\mathrm{fit}} + \lambda_p \mathcal{E}(P_I, P_T) + \lambda_g(e) \mathcal{E}\left(\mathcal{N}(P_I)\mathcal{N}(P_T)^\top, S_{\mathcal{B}}^e \right).$$

The separately constructed graph-ranking objective is

$$\mathcal{L}_{\mathrm{RANK}} = \mathcal{R}_{\mathrm{rank}}\left(P_I, P_T; S_{\mathcal{B}}^e \right),$$

yielding

$$\mathcal{L}_{\mathrm{hash}} = \mathcal{L}_{\mathrm{DISTILL}} + \lambda_r \mathcal{L}_{\mathrm{RANK}}.$$

GJHL is optimized before the modality-specific hash functions within each epoch, and their outputs affect any scheduled similarity-matrix update only afterward.

At inference time, only f_I and f_T are retained:

$$b_q^I = \operatorname{sign}\left(f_I(\mathcal{N}(x_q)) \right), \quad b_q^T = \operatorname{sign}\left(f_T(\mathcal{N}(y_q)) \right).$$

Thus, CSCI, HASR, and GJHL are used only during training, while each query is encoded independently by its corresponding modality-specific hash function.

C Additional Experimental Results

We provide additional experiments to further characterize CSCIH beyond the aggregate retrieval performance reported in the main paper. Specifically, we examine retrieval behavior across different operating ranges, conduct fine-grained ablations and progressive build-up experiments, present a qualitative case study of context-induced retrieval errors, and analyze the computational cost and graph-free retrieval efficiency of CSCIH. Together, these results provide a more detailed view of the contributions of the proposed components, the retrieval behavior of the resulting representations, and the computational characteristics of the overall framework.

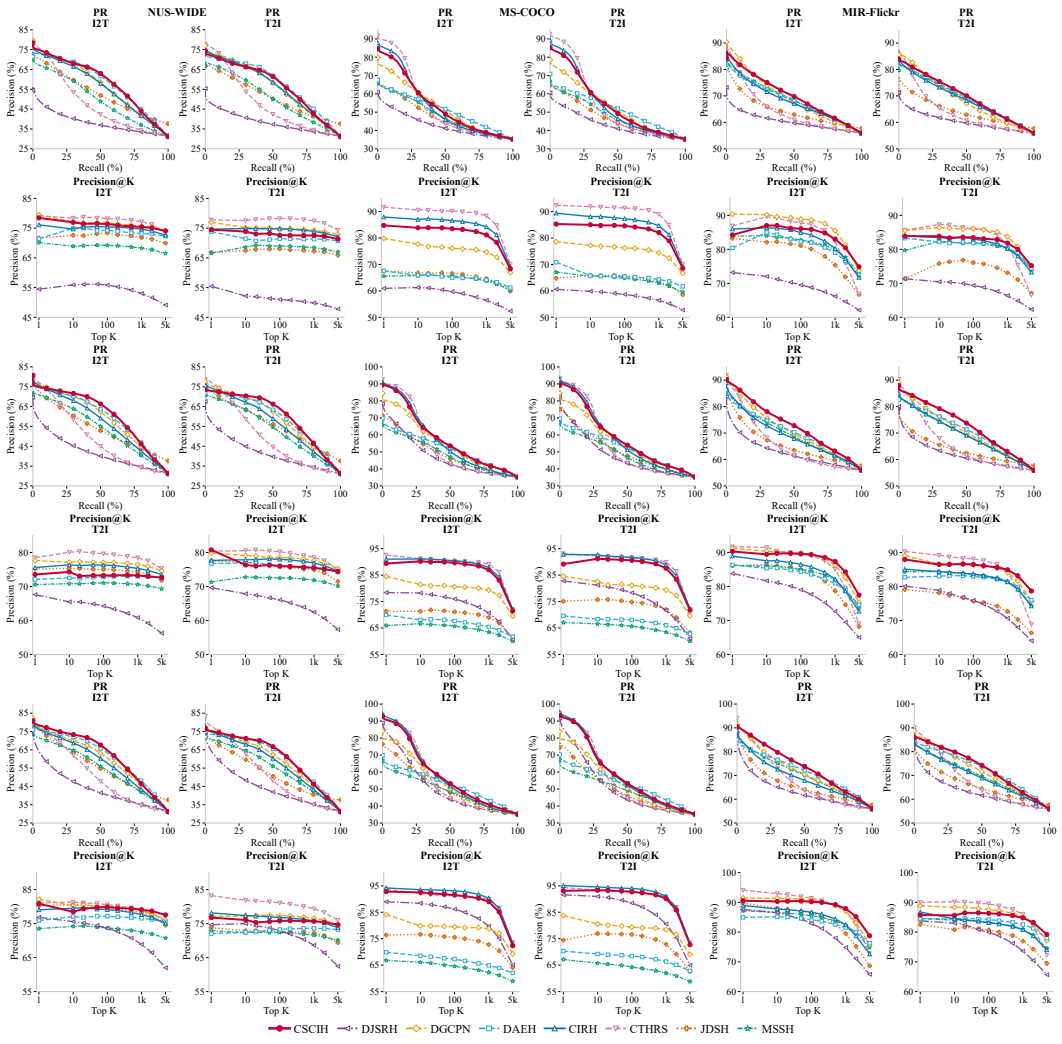


Figure 4: Additional retrieval results on MIRFLICKR-25K, NUS-WIDE, and MS-COCO using 16-, 32-, and 64-bit hash codes. For each hash-code length, rows show precision-recall and Precision@K curves, with I2T and T2I shown for each dataset.

C.1 Additional Hash Retrieval Results

To complement the aggregate mAP results reported in the main paper, we further examine retrieval behavior across different operating ranges and hash-code lengths. Fig. 4 reports precision-recall and Precision@K curves on all three benchmarks using 16-, 32-, and 64-bit hash codes in both retrieval directions. These results characterize ranking behavior across different recall levels and retrieval depths rather than summarizing the entire ranking with a single score.

Across MIRFLICKR-25K and NUS-WIDE, CSCIH maintains favorable retrieval performance over most operating ranges and across the three hash-code lengths. On MS-COCO, several competing methods remain competitive at particular operating points, especially at shallow retrieval depths, while CSCIH maintains strong performance as the retrieval range expands. Overall, these trends are consistent with the mAP results in the main paper and show that the gains of CSCIH are not confined to a particular retrieval direction, hash-code length, or narrow operating region, but remain evident across a broad range of recall levels and retrieval depths.

Table 3: Fine-grained ablation study of CSCIH (%) on MIRFLICKR-25K, NUS-WIDE, and MS-COCO.

Variant	CSCI	HASR	GJHL	MIRFLICKR-25K		NUS-WIDE		MS-COCO	
				I→T	T→I	I→T	T→I	I→T	T→I
w/o Fixed Split	○	√	√	72.7	73.6	61.8	62.9	63.7	63.6
w/o Low-Rank Perturbation	○	√	√	72.2	73.1	60.3	61.5	60.5	60.4
w/o Fixed Fusion	√	○	√	74.1	74.3	65.1	65.6	64.3	64.0
w/o Hash-only Refresh	√	○	√	73.7	74.6	65.1	65.7	64.1	63.9
w/o Periodic Refresh	√	○	√	73.9	74.6	64.3	64.8	63.7	63.3
w/o Vanilla GCN	√	√	○	73.5	74.5	64.7	65.0	62.9	62.7
w/o Pairwise Cosine	√	√	○	73.6	74.5	64.7	65.7	62.9	62.8
w/o Static Cross-Graph	√	√	○	73.6	74.1	63.4	63.8	63.5	63.1
Full CSCIH	√	√	√	**74.1**	**74.9**	**65.3**	**66.2**	**64.5**	**64.7**

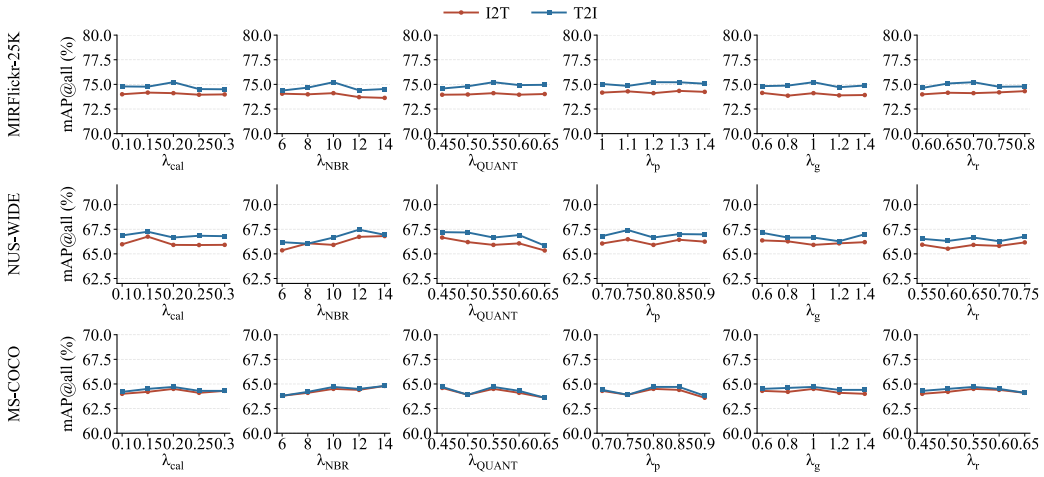


Figure 5: Sensitivity of CSCIH to six representative hyperparameters on MIRFlickr-25K, NUS-WIDE, and MS-COCO with 32-bit hash codes.

C.2 Additional Ablation Experiments

To further examine the roles of the main designs in CSCIH, we conduct fine-grained subtractive ablations by replacing each component with simpler alternatives while keeping the remaining components unchanged. As shown in Table 3, the full CSCIH achieves the best or tied-best performance across all settings, while none of the replacement variants consistently recovers its performance.

For CSCI, both the fixed split and low-rank perturbation degrade performance, indicating that the gain cannot be recovered by a fixed feature partition or generic perturbation alone, but benefits from learning subject and context oriented representations under contextual variation. For HASR, fixed fusion, hash-only refresh, and periodic refresh fail to consistently recover the full performance, supporting the adaptive balance between hash- and subject derived similarity relations according to subject consistency and training progress. For GJHL, replacing the proposed design with a vanilla GCN, pairwise cosine relations, or a static cross-modal graph also reduces performance, supporting the organization of modality-local and cross-modal relational structures through complementary Euclidean and hyperbolic geometries.

Overall, these results support the complementary roles of the three components: CSCI provides subject oriented information whose consistency is assessed across alternative contextual states, HASR regulates its contribution to subsequent similarity updates, and GJHL organizes the resulting relational information to guide modality-specific hash learning.

C.3 Additional Parameter Analysis

To complement the parameter analysis in the main paper, we further evaluate the sensitivity of CSCIH across all three benchmark datasets. Fig. 5 reports the results for λ_{cal}, λ_{NBR}, λ_{QUANT}, λ_p, λ_g, and λ_r using 32-bit hash codes. These parameters respectively control CSCI calibration, neighborhood reconstruction and quantization in the teacher-side objective, and paired cross-modal consistency, similarity-relation preservation, and graph-ranking supervision in modality-specific hash learning.

Across MIRFlickr-25K, NUS-WIDE, and MS-COCO, retrieval performance remains relatively stable over the evaluated ranges, with I2T and T2I exhibiting closely aligned trends. MIRFlickr-25K and MS-COCO show only modest variations across the six parameters, while NUS-WIDE is somewhat more responsive to λ_{NBR} and λ_{QUANT}, without exhibiting substantial performance degradation. These results are consistent with the observations in the main paper and further show that the limited sensitivity of CSCIH is maintained across datasets and retrieval directions.

C.4 Computational Complexity

Let N denote the number of training pairs. Constructing the initial similarity matrices requires pairwise relation estimation over the training set and therefore incurs $O(N^2)$ complexity. Within each epoch, CSCI and GJHL operate on mini-batches and local neighborhoods, giving $O(N)$ complexity with respect to N when the model and training configurations are fixed. At a scheduled similarity-matrix update, HASR constructs the hash-derived and subject derived similarity relations over the training set, resulting in $O(N^2)$ complexity.

Therefore, with respect to N, the overall training complexity is dominated by the scheduled $O(N^2)$ similarity-matrix updates rather than the per-epoch $O(N)$ optimization. Since these updates are performed only at scheduled epochs, the quadratic computation is not incurred at every optimization step. In practice, chunked matrix computation and sparse neighborhood propagation reduce peak memory usage and runtime overhead without changing the asymptotic complexity.

C.5 Training and Retrieval Efficiency

Table 4 reports the training and retrieval efficiency of CSCIH and representative baselines on NUS-WIDE with 32-bit hash codes. CSCIH requires 9.979 M parameters during training and 2.413 s per epoch, reflecting the additional computation required by CSCI, HASR, and GJHL. Despite this increase in training computation, CSCIH uses only 0.316 GB peak GPU memory, remaining below CIRH, DGCPN, and CTHRS under the evaluated configurations.

At inference time, only the modality-specific hash functions f_I and f_T are retained, reducing the parameter count from 9.979 M to 4.506 M, while CSCI, HASR, and GJHL are used only during training. Each query is therefore encoded independently using its corresponding modality-specific hash function, with encoding time below 0.001 ms in both retrieval directions. Overall, these results show that the additional subject-context modeling and relational organization are concentrated in training, without introducing the corresponding training-time structures into retrieval.

Table 4: Training and retrieval efficiency on NUS-WIDE with 32-bit hash codes.

Methods	Parameters (M)		Training		Query Encoding (ms)	
	Train	Infer.	VRAM (GB)	Time/Ep. (s)	I→T	T→I
DJSRH	4.465	4.465	0.122	0.452	0.002	0.002
DGCPN	1.116	1.116	1.306	1.154	0.002	0.001
CIRH	4.478	2.511	0.599	0.403	0.001	< 0.001
MSSH	4.532	4.532	0.078	0.501	0.076	0.075
CTHRS	21.264	21.264	1.318	0.695	0.003	0.003
CSCIH	9.979	4.506	0.316	2.413	< 0.001	< 0.001

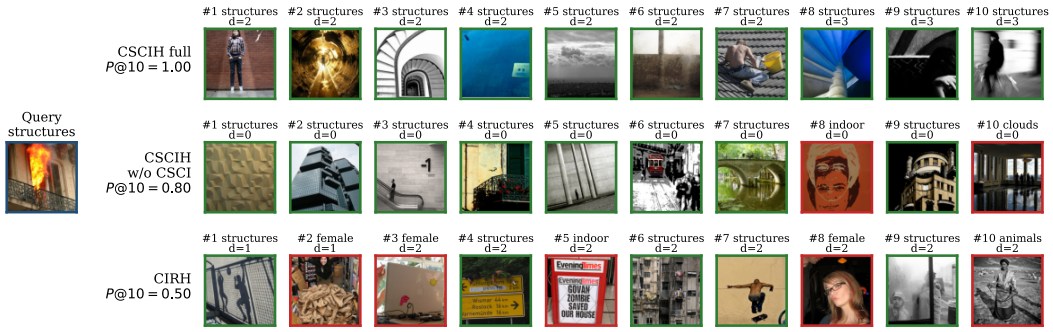


Figure 6: Qualitative comparison on MIRFlickr-25K for image-to-text retrieval with 32-bit hash codes. Green and red borders denote relevant and irrelevant results, respectively. Each retrieved text instance is visualized using its paired image.

C.6 Case Study

To complement the aggregate retrieval results, we further compare CIRH, CSCIH without CSCI (w/o CSCI), and the full CSCIH model on a representative image-to-text query from MIRFlickr-25K. For visualization, each retrieved text instance is represented by its paired image. As shown in Fig. 6, the query belongs to the *structures* category. CIRH retrieves five relevant instances within the top ten, yielding $\mathrm{P@10} = 0.50$, while several irrelevant samples occupy high-ranking positions. The w/o CSCI variant improves the ranking to $\mathrm{P@10} = 0.80$, but irrelevant results remain. In contrast, the full CSCIH retrieves relevant instances at all ten positions and achieves $\mathrm{P@10} = 1.00$, despite substantial variation in the visual appearance of the retrieved samples.

The progression from CIRH to the CSCI-removed variant and then to the full model provides a qualitative counterpart to the quantitative ablation results. In particular, the further improvement obtained by the full model is consistent with the role of CSCI in providing subject oriented information whose consistency is assessed across alternative contextual states before contributing to subsequent similarity updates. The example therefore illustrates how the additional subject oriented information introduced by CSCI is reflected in the final retrieval ranking, with fewer irrelevant samples appearing among the highest-ranked results.

D AI Use Statement

Generative AI tools were used solely for language polishing to improve the clarity and readability of the manuscript. The authors reviewed all AI-assisted edits and take full responsibility for the final content of the paper.