SING-CH: Task-Scale-Agnostic Lifelong Cross-Modal Hashing on Statistical Manifolds

Haoran Yang[1] **Junge Chen**[2] **Jun Long**[1] **Zhan Yang**[1*]

[1]Central South University, China
[2]Xiamen University, China
{haoran.yang, junlong, zyang22}@csu.edu.cn, chenjunge@stu.xmu.edu.cn

Abstract

Lifelong cross-modal hashing demands incremental updates as new tasks arrive, yet existing regularization-based methods such as EWC and SI share a weakness: their effective regularization strength is tied to task scale, because importance is accumulated over supervision units whose count grows with the task. This forces costly per-task hyperparameter tuning in streaming deployments. We propose SING-CH, a lifelong hashing framework built on information geometry. We recast the pairwise objective as a per-pair Gibbs likelihood, which fixes the statistical unit to one cross-modal relation and yields a per-pair $1/n_t^2$ Fisher information. We prove this estimator is exactly invariant to dataset replication, measure that its scale varies by only $1.31\times$ across a $1{,}192\times$ range of task sizes, and measure that the shape a preconditioner consumes varies by $9.7\times$. A metric-mismatch theorem shows that applying natural gradient to the composite lifelong objective reintroduces size dependence, which makes the decoupled design, the FIM as regularizer with Adam as the preconditioner, a structural necessity. On three benchmarks, SING-CH outperforms 13 baselines by 4.1 mAP points over the strongest one without per-task tuning, and under extreme task-scale imbalance it degrades by 1.6 points (2.0%) versus 2.9 points (3.9%) for the strongest baseline.

1 Introduction

Cross-modal retrieval has become an indispensable capability in modern information systems [1]: given a query from one modality, the goal is to efficiently retrieve semantically relevant items from another modality. Hashing-based methods have emerged as the dominant paradigm for large-scale cross-modal retrieval, encoding high-dimensional multimodal features into compact binary codes that support sub-linear search via Hamming distance [2]. Deep cross-modal hashing methods, including Deep Cross-Modal Hashing (DCMH) [3], Deep-Joint-Semantics Relevance Hashing (DJSRH) [4], and Joint-modal Distribution-based Similarity Hashing (JDSH) [5], have demonstrated strong retrieval performance by jointly learning modality-specific hash functions through shared semantic supervision. Despite these advances, existing methods universally assume a static, fully available training corpus. This assumption is violated in real-world deployments where data arrives as a continuous stream of tasks.

In practice, cross-modal retrieval systems must operate under a lifelong learning regime: new tasks arrive sequentially covering different semantic categories, and the system must incorporate new knowledge without forgetting previously learned representations. This challenge, known as catastrophic forgetting [6], is particularly acute in cross-modal hashing because the binary code space is highly constrained. Regularization-based continual learning methods such as Elastic Weight

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Consolidation (EWC) [7] and Synaptic Intelligence (SI) [8] address forgetting by penalizing changes to parameters deemed important for previous tasks, using the FIM as a measure of parameter importance. However, these methods suffer from an underappreciated limitation: their regularization strength is entangled with task scale.

Task scale enters through a unit mismatch. Importance-based methods protect the example, but the supervision unit of a pairwise objective is the relation, and a task contributes n_t^2 of them; an importance estimator built on the example therefore inherits the size of the task rather than the size of a relation. With a fixed λ the effective strength of the penalty follows task composition: when one task holds 500 pairs and another 50,000, it fluctuates by orders of magnitude. This forces practitioners to either tune λ per task or accept severe degradation. These methods were also designed for unimodal classification objectives, and pairwise similarity objectives do not admit a natural probabilistic interpretation, which is what a likelihood-based fix needs.

We address these limitations through a principled information-geometric reformulation. The key insight is that the empirical FIM of the pairwise objective becomes a $1/n_t^2$ per-pair average, which is what makes it replication-invariant. To make this framework applicable to cross-modal hashing, we introduce a probabilistic reconstruction of the pairwise hashing objective: we model the pairwise label S_{ij} as a Gibbs distribution over the squared error between S_{ij} and the relaxed similarity $Q_{ij}(\theta)$, motivated by the information-geometric structure of exponential families [9]. This reconstruction gives the parameter space a statistical manifold structure, enabling the FIM to be analytically computed via the Straight-Through Estimator [10].

We propose Scale-Invariant Normalized Gradient for Cross-modal Hashing (SING-CH), a lifelong framework operating on statistical manifolds. SING-CH employs the scale-normalized FIM as a geometric regularizer to protect old-task knowledge, while delegating gradient preconditioning to Adam. This decoupled design is motivated by the recently established equivalence between Adam and diagonal natural gradient [11], and by a theoretical analysis (Section 3.5) showing that directly applying natural gradient to the composite lifelong objective reintroduces scale dependence.

Our contributions are:

1. **Pairwise Gibbs normalization** (Section 3.2): We recast cross-modal hashing as a per-pair maximum likelihood, fixing the statistical unit to one relation and yielding the $1/n_t^2$ per-pair Fisher of Eq. (4).
2. **Replication-invariant importance** (Theorem 3.4): We prove the per-pair FIM is exactly invariant to dataset replication, give a diagonal-approximation bound that depends on spectral quantities only, and measure that its scale varies by $1.31\times$ across a $1{,}192\times$ range of task sizes (Section 4.6).
3. **Optimal regularization without per-task tuning** (Theorems 3.8–3.9): We characterize the forgetting-rigidity trade-off and show the optimal λ^* depends on spectral properties of the FIM but not on individual task sizes.
4. **Why natural gradient fails here** (Theorem 3.7): We show that preconditioning the composite objective by the current task's Fisher applies a metric-mismatch operator to the penalty whose norm depends on the current task's size.

2 Related Work

Cross-modal hashing techniques map heterogeneous inputs into a shared Hamming space to preserve semantic similarity. Kumar and Udupa [12] proposed CVH and Song et al. [13] developed IMH to exploit inter-modal correlation for unsupervised retrieval. To enhance retrieval quality through supervision, Jiang and Li [3] introduced DCMH, while subsequent researchers developed DJSRH [4] and JDSH [5]. More recently, Jin et al. [14] designed graph attention networks, Zhang et al. [15] implemented dynamic masking, Xia et al. [16] utilized CLIP-based hashing, and both He et al. [17] and Huang et al. [18] proposed contrastive approaches to optimize feature alignment. While these methods assume a static training environment, SING-CH enables incremental updates by specifically addressing the challenge of catastrophic forgetting.

Various researchers have explored continual learning through distinct mechanisms to preserve knowledge. Kirkpatrick et al. [7] proposed EWC, Zenke et al. [8] introduced SI, and Chaudhry et al. [19]

developed RWalk to penalize changes to important parameters. In the replay-based setting, Lopez-Paz and Ranzato [20] established GEM, while others implemented A-GEM [21] and iCaRL [22] to utilize replay-based strategies. Mallya and Lazebnik [23] proposed PackNet and Wang et al. [24] devised L2P to adapt network architectures for sequential tasks. From an information-geometric perspective, Ritter et al. [25] introduced the online Kronecker-factored Laplace approximation, and Nguyen et al. [26] framed the learning process as online Bayesian inference through VCL. SING-CH instead removes the scale dependence at its source and states the invariance it guarantees explicitly.

In lifelong hashing, Xu et al. [27] developed DLCMH as a direct predecessor to modern frameworks. Zhan et al. [28] proposed POLISH, Chen et al. [29] introduced DCGH, and Li et al. [30] designed LCDH to advance incremental retrieval, yet these works do not employ information-geometric frameworks or provide theoretical guarantees regarding task-scale variation. The natural gradient established by Amari [31] and the K-FAC method proposed by Martens and Grosse [32] form the technical foundation of our approach. Recently, Hwang et al. [11] proved that Adam implicitly performs diagonal natural gradient descent using the empirical Fisher, which justifies the decoupled design of SING-CH where Adam handles new-task preconditioning and the explicit FIM ensures old-task protection. We further prove in Section 3.5 that this decoupling is a structural necessity, as directly applying the natural gradient to composite objectives reintroduces scale dependence.

3 Method

3.1 Problem Formulation

We consider a lifelong cross-modal hashing setting where a model learns sequentially from a task stream $\mathcal{T} = \{T_1, T_2, \ldots, T_K\}$. Each task T_t provides n_t image-text pairs $\{(v_i, w_i)\}_{i=1}^{n_t}$ with pairwise semantic similarity labels $S_{ij} \in \{0, 1\}$. Once task T_t is completed, its training data becomes inaccessible. Task scales $\{n_t\}$ may vary by orders of magnitude.

The primary objective of this work is to learn hash functions $f^v : \mathbb{R}^{d_v} \to \{-1, +1\}^L$ and $f^w : \mathbb{R}^{d_w} \to \{-1, +1\}^L$ that preserve cross-modal semantic similarity via Hamming distance, retain performance on all previous tasks without data replay, and remain robust to extreme task-scale imbalance without the need for per-task hyperparameter tuning.

Definition 3.1 (Task-Scale-Agnostic)**.** *A lifelong learning method is* task-scale-agnostic *if its parameter update direction is invariant under uniform dataset replication: replacing n_t training samples with αn_t identical copies ($\alpha \in \mathbb{Z}^+$) leaves the update $\Delta\theta$ unchanged.*

Notation. Model parameters $\theta = \{\theta^v_{\text{enc}}, W^v, c^v, \theta^w_{\text{enc}}, W^w, c^w\} \in \mathbb{R}^P$. $\hat{F}^{(t)}$ denotes the empirical FIM of task t, $F^{(t)}_{\text{old}} = \sum_{s=1}^{t} \hat{F}^{(s)}$ the cumulative FIM, and θ_t the parameters after completing task t.

3.2 Probabilistic Cross-Modal Hashing Model

Figure 1 illustrates the overall architecture of SING-CH, which integrates pairwise Gibbs normalization with Fisher-weighted manifold regularization to achieve scale-agnostic lifelong cross-modal hashing.

Dual-Modal Encoders. The image encoder maps raw images to continuous hash embeddings:

$$\tilde{b}^v_i = \tanh(\beta \cdot W^v \phi^v(v_i; \theta^v_{\text{enc}}) + c^v)\,, \tag{1}$$

where ϕ^v is a backbone network, $W^v \in \mathbb{R}^{L \times d_{\text{feat}}}$ is the hash projection, and $\beta > 0$ is a temperature parameter annealed from β_0 to $\beta_{\max}$ within each task. Discrete hash codes are $b^v_i = \text{sign}(\tilde{b}^v_i)$; the text encoder follows the same structure.

Gibbs Probabilistic Model. The cross-modal similarity score is $Q_{ij}(\theta) = \frac{1}{2}(\tilde{b}^v_i \cdot \tilde{b}^w_j + \tilde{b}^v_j \cdot \tilde{b}^w_i)$. We model the probability of observing S_{ij} as a Gibbs distribution:

$$P(S_{ij} \mid \theta) = \frac{1}{Z_{ij}} \exp\big(-\|S_{ij} - Q_{ij}(\theta)\|^2\big)\,, \quad Z_{ij} = \exp(-Q^2_{ij}) + \exp(-(1 - Q_{ij})^2). \tag{2}$$

The Gibbs form endows the parameter space with a Riemannian metric structure: $\frac{1}{2}\Delta\theta^\top F \Delta\theta$ is the second-order expansion of $D_{\text{KL}}(P(\cdot|\theta_t) \| P(\cdot|\theta))$; the pairwise structure supplies the $1/n_t^2$ per-pair

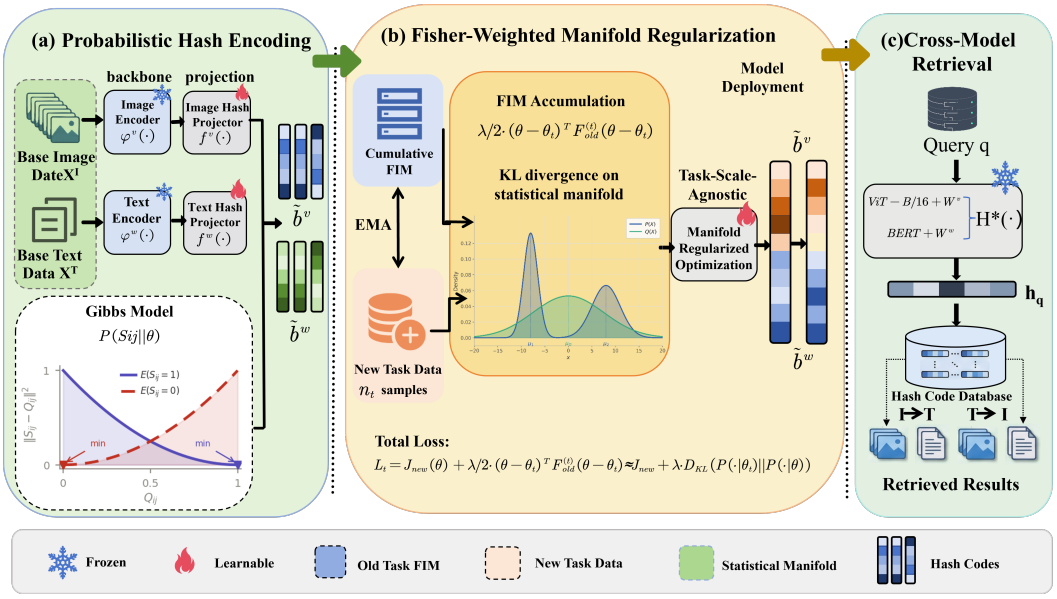

Figure 1: Overview of the proposed SING-CH framework.

normalization of Eq. (4); and the probabilistic framework makes approximation errors analytically tractable. The training loss is:

$$J_{\text{new}}(\theta) = \frac{1}{n_t^2} \sum_{i,j=1}^{n_t} \left[\|S_{ij} - Q_{ij}(\theta)\|^2 + \log Z_{ij}(\theta) \right] = -\frac{1}{n_t^2} \log \prod_{i,j} P(S_{ij} \mid \theta). \tag{3}$$

This is the negative log-likelihood *per ordered pair*: the $1/n_t^2$ factor turns n_t^2 per-pair terms into an average. Fixing this statistical unit, one cross-modal relation rather than one task, is what makes the estimator replication-invariant in the sense of Definition 3.1.

Lemma 3.2 (Score-approximation error). *The score decomposes as* $\nabla_\theta \log P = 2(S_{ij} - Q_{ij}) \nabla_\theta Q_{ij} - \nabla_\theta \log Z_{ij}$, *and the empirical Fisher retains the first term. The omitted term satisfies* $\|\nabla_\theta \log Z_{ij}\| \le g(Q_{ij}) \cdot \|\nabla_\theta Q_{ij}\|$ *with* $g(q) \le 2|2q-1|$, $g(1/2) = 0$, *and* $\|\nabla_\theta \log Z_{ij}\| \to 0$ *as* $\beta \to \infty$.

3.3 Fisher Information on Statistical Manifolds

The empirical FIM retains the residual term of the score:

$$\hat{F}^{(t)}(\theta) = \frac{4}{n_t^2} \sum_{i,j=1}^{n_t} (S_{ij} - Q_{ij}(\theta))^2 \, \nabla_\theta Q_{ij}(\theta) \, \nabla_\theta Q_{ij}(\theta)^\top. \tag{4}$$

Dividing by n_t^2 turns the n_t^2-term sum of per-pair Fisher matrices into a per-pair average. That average is the object we protect: its *scale* is invariant to task size, while its *shape* is subject to sampling noise. Theorem 3.7 rests on this distinction.

Lemma 3.3 (Scale-Invariance). *Under dataset replication* ($n_t \to \alpha n_t$), *the empirical FIM remains unchanged: the numerator grows by* α^2 *while the denominator also grows by* α^2. *With objective and Fisher both mean-normalized (Eqs.* (3)–(4)*), the same cancellation applies to* ∇J_{new}, *so* $\Delta\theta^{(\alpha n_t)} = \Delta\theta^{(n_t)}$.

Theorem 3.4 (Replication invariance of the per-pair FIM). *Let* $\hat{F}(\theta)$ *be the per-pair empirical Fisher of Eq.* (4) *at a fixed* θ, *and* $\hat{F}_\alpha$ *the same estimator computed after replacing a task by* α *identical copies of itself (Definition 3.1). Then* $\hat{F}_\alpha = \hat{F}$ *exactly, so* $\lambda_{\min}$, $\lambda_{\max}$, *and* κ_F *are invariant, and the importance weights in* $\frac{\lambda}{2} \sum_p F_{\text{old},p}(\theta_p - \theta_{\text{prev},p})^2$ *do not drift with replication. With* $D = \text{diag}(\hat{F})$,

$\|\Delta\theta_{\text{diag}} - \Delta\theta_{\text{exact}}\| \leq \frac{\kappa_F - 1}{\lambda_{\min}(\hat{F})} \|\nabla J\|$, *a bound that depends on the spectrum of* $\hat{F}$ *only. The invariance is exact under replication; Section 4.6 reports how independently sampled tasks of different sizes perturb the spectrum.*

Remark 3.5 (Kronecker approximations). The replication argument carries over to K-FAC factors $\tilde{A}_l \otimes \tilde{G}_l$ with fixed damping ε, since replication leaves A_l, G_l unchanged. The implementation uses the diagonal alone, measured directly in Section 4.6.

3.4 The SING-CH Algorithm

Operationally the method has three steps, and only the third touches the optimizer: (i) estimate the diagonal importance of each cross-modal relation from the Gibbs score of Eq. (4); (ii) calibrate each task's importance to unit mean and accumulate it into F_{old}; (iii) add the penalty $\frac{\lambda}{2} \sum_p F_{\text{old},p}(\theta_p - \theta_{\text{prev},p})^2$ to an otherwise unchanged Adam objective. The implementation keeps the diagonal alone.

The SING-CH training objective combines a Gibbs-derived pairwise loss with auxiliary terms that improve practical discriminability without participating in FIM estimation:

$$J(\theta) = \underbrace{w_g \cdot J_{\text{gibbs}}(\theta)}_{\text{FIM source}} + \underbrace{w_p \cdot J_{\text{proxy}}(\theta) + w_v \cdot J_{\text{var}}(\theta) + w_d \cdot J_{\text{decorr}}(\theta)}_{\text{auxiliary (not in FIM)}} + \frac{\lambda}{2} \Delta\theta^\top F_{\text{old}}^{(t)} \Delta\theta, \quad (5)$$

where $F_{\text{old}}^{(t)} = \sum_{s=1}^{t-1} \hat{F}^{(s)}$. The Gibbs pairwise loss J_{gibbs} (Eq. (3)) provides the theoretical foundation: the $1/n_t^2$ FIM normalization and all theorems derive from this component. The auxiliary terms serve as practical enhancements: a proxy loss (w_p=0.6) projects hash codes toward learnable class centers for intra-class compactness; a variance loss (w_v) reduces combinatorial semantic drift in multi-label datasets such as MS-COCO; and a decorrelation loss (w_d=0.05) encourages independent hash bits to prevent code collapse. The FIM is computed from J_{gibbs} alone via Eq. (4), which preserves the scale-agnostic property for all theoretical guarantees. The auxiliary losses shape the optimization landscape without affecting the Fisher-Rao geometry.

Remark 3.6. Unlike standard EWC, the $1/n_t^2$ per-pair normalization follows from taking the relation, not the task, as the statistical unit. The FIM serves as a *geometric constraint*: $\frac{1}{2}\Delta\theta^\top F_{\text{old}} \Delta\theta$ is the second-order expansion of D_{KL} on the statistical manifold, so new task learning deviates minimally from previous optimal solutions in the Fisher-Rao geometry. Adam handles preconditioning; Hwang [11] proved that Adam implicitly performs diagonal natural gradient descent.

3.5 Why Not Natural Gradient Optimization?

A natural question is whether SING-CH should directly use natural gradient to optimize the composite objective, replacing Adam with $\theta \leftarrow \theta - \eta \tilde{F}^{-1} \nabla J$ in Algorithm 1. We show this is theoretically unsound.

Theorem 3.7 (Metric mismatch of composite natural-gradient updates)**.** *Let* $\hat{F}^{(t)}$ *be the per-pair Fisher of the current task, estimated from* n_t^2 *pairs, and* F_{old} *the accumulated Fisher of completed tasks. Preconditioning the composite gradient by* $[\hat{F}^{(t)}]^{-1}$ *applies the operator* $[\hat{F}^{(t)}]^{-1} F_{\text{old}}$ *to the penalty, so the effective regularization strength is* $\lambda \|[\hat{F}^{(t)}]^{-1} F_{\text{old}}\|$ *rather than* λ*; this operator is the identity only when current and past tasks induce the same metric. Because* $\hat{F}^{(t)}$ *is an average over* n_t^2 *pairs, its entries concentrate at rate* $O(1/n_t)$*, so its small eigenvalues become noise-dominated as* n_t *shrinks and the operator norm grows. A single Fisher preconditioner therefore cannot hold the data and penalty terms in a fixed ratio across task sizes.*

Proof Sketch. The data term is invariant: ∇J_{new} and $\hat{F}^{(t)}$ are per-pair means over the same n_t^2 pairs (Eqs. (3)–(4)), so $[\hat{F}^{(t)}]^{-1} \nabla J_{\text{new}}$ does not depend on n_t. The penalty term is not: F_{old} is fixed by completed tasks while $[\hat{F}^{(t)}]^{-1}$ is the inverse of a sample average whose conditioning depends on n_t. The cancellation that makes the natural gradient scale-free within one model cannot act across the composite objective. Full argument in Appendix B.6. □

On composite objectives with heterogeneous gradient sources, no single Fisher preconditioner normalizes both components. The decoupled design of SING-CH, which uses the FIM as a geometric

Algorithm 1 SING-CH Training Procedure

Input: Task stream $\mathcal{T} = \{T_1, \ldots, T_K\}$, λ, η, $[\beta_0, \beta_{\max}]$, ε
Output: Updated parameters θ_K

Initialize θ_0, $F_{\text{old}} \leftarrow \mathbf{0}$
for $t = 1, \ldots, K$ **do**
  Store $\theta_{\text{prev}} \leftarrow \theta_{t-1}$, reset $\beta \leftarrow \beta_0$
  for epoch $= 1, \ldots, E$ **do**
    Anneal β from β_0 to $\beta_{\max}$
    for each mini-batch $\mathcal{B} \subset T_t$ **do**
      $J_{\text{gibbs}} = \frac{1}{|\mathcal{B}|^2} \sum_{(i,j)\in\mathcal{B}} [\|S_{ij} - Q_{ij}\|^2 + \log Z_{ij}]$ {FIM source}
      $J_{\text{aux}} = w_p J_{\text{proxy}} + w_v J_{\text{var}} + w_d J_{\text{decorr}}$ {not in FIM}
      $J_{\text{reg}} = \frac{\lambda}{2} \sum_p F_{\text{old},p} (\theta_p - \theta_{\text{prev},p})^2$
      $\theta \leftarrow \theta - \eta \cdot \text{Adam}(\nabla(J_{\text{gibbs}} + J_{\text{aux}} + J_{\text{reg}}))$
    end for
  end for
  $\hat{F}^{(t)} = \frac{4}{|\mathcal{B}|^2} \sum_{i,j} (S_{ij} - Q_{ij})^2 \nabla Q_{ij} \nabla Q_{ij}^\top$
  $F_{\text{old}} \leftarrow F_{\text{old}} + \hat{F}^{(t)}$, $\theta_t \leftarrow \theta$
end for
return θ_K

regularizer for old tasks and Adam as an implicit diagonal natural gradient for new-task optimization, follows as a theoretical consequence.

Theorem 3.8 (Cumulative Forgetting Bound)**.** *$J_t(\theta_K) - J_t(\theta_t) \leq \frac{L_J (K-t)^2 G_{\max}^2}{2\lambda^2 \lambda_F^2}$, given L_J-smoothness of J_t, $\|\nabla J_s\| \leq G_{\max}$, and $F_{\text{old}} \succeq \lambda_F I$.*

Theorem 3.9 (Forgetting-Rigidity Trade-off)**.** *With $\mathcal{F}(\lambda) \leq C_F/\lambda^2$ and $\mathcal{R}(\lambda) \leq C_R \lambda^2/\mu_{\text{PL}}$, the cost $\mathcal{F} + \mathcal{R}$ is minimized at $\lambda^* = (C_F \mu_{\text{PL}}/C_R)^{1/4}$, which contains no explicit n_t.*

4 Experiments

4.1 Experimental Setup

We evaluate on three benchmarks: MIRFlickr-25K [33] (20015 pairs, 24 classes), NUS-WIDE TC21 [34] (50000 pairs, 21 classes), and MS-COCO [35] (50000 pairs, 80 classes). All use VGG-4096 image features and BoW text features. We adopt a training-exclusive retrieval pool to prevent metric inflation through memorization.

Task Construction. Three protocols:

- **Uniform** (K=5): even class split, ∼4000 pairs/task on MIRFlickr and ∼10000 on NUS-WIDE and MS-COCO.
- **Extreme** (K=5): imbalanced split; task sizes span 29–34562 pairs (1192×).
- **Long** (K=10): ∼2000 pairs/task on MIRFlickr and ∼5,000 on the 50K benchmarks.

SING-CH results averaged over 4 random seeds; all other methods single seed.

We compare against 13 baselines: regularization-based CL (EWC [7], NormalizedEWC, DER++ [36]), lifelong hashing (DLCMH [27], POLISH [28], LCDH [30], GAH [14], PromptHash [37], DCGH [29]), and static hashing (DCH [38], DCMH [3], SCM [39], SePH [40]).

The implementation utilizes a VGG-4096 encoder and projection layers ($4096 \to 4096 \to L$), while the BoW text encoder follows a ($1386 \to 4096 \to L$) architecture. Optimization uses Adam ($\eta = 10^{-4}$, hash projection 10^{-3}) with $L = 64$ bits, a batch size of 64, and 30 epochs per task. Hyperparameters λ, β_0, $\beta_{\max}$, and w_v are set once per dataset from its statistics such as sample size, class cardinality, and label sparsity, and held fixed across the uniform, extreme, and long protocols. Specifically, $\lambda \in \{5, 10, 20\}$, $\beta_0 \in \{0.3, 0.5, 1.0\}$, $\beta_{\max} \in \{8, 10\}$, and $w_v \in \{1.0, 1.2, 2.5\}$ are assigned to MIRFlickr, NUS-WIDE, and MS-COCO respectively, with no per-task tuning. Following

Table 1: Lifelong cross-modal retrieval (mAP@L=64, 5-task uniform). **Bold**: best. Underline: second. SING-CH: 4-seed avg ($\sigma < 0.003$). All methods use VGG+BoW features under identical lifelong protocol.

Method	Source	MIRFlickr	NUS-WIDE	MS-COCO	Avg
Non-continual hashing					
DCMH	CVPR'17	0.577	0.405	0.422	0.468
SCM	AAAI'14	0.664	0.427	0.660	0.584
SePH	CVPR'15	0.694	0.447	0.614	0.585
DCH	CVPR'18	0.614	0.429	0.435	0.493
Lifelong hashing					
DLCMH	TCSVT'24	0.568	0.386	0.634	0.529
POLISH	WWW'24	0.562	0.390	0.535	0.495
LCDH	AAAI'25	0.659	0.454	0.572	0.562
GAH	NeurIPS'24	0.661	0.444	0.631	0.579
PromptHash	CVPR'25	0.566	0.398	0.540	0.501
DCGH	Appl.Sci.'25	<u>0.799</u>	<u>0.630</u>	<u>0.793</u>	<u>0.740</u>
Regularization-based CL					
EWC	PNAS'17	0.552	0.443	0.530	0.508
NormalizedEWC	—	0.567	0.421	0.611	0.533
DER++	NeurIPS'20	0.570	0.399	0.535	0.502
SING-CH	—	**0.799**	**0.745**	**0.800**	**0.781**

Section 3.5, Adam handles new-task preconditioning while the explicit FIM term protects old tasks, which avoids the scale re-dependence identified in Theorem 3.7.

State and cost. Consolidation keeps exactly two $|\theta|$-sized float32 vectors, the task-end snapshot and the accumulated diagonal importance, so persistent state is $8|\theta|$ bytes regardless of how many tasks have been completed. The boundary pass costs $O(B^2L)$ for batch size B, linear in the current task. End to end, Full and a no-Fisher control take $1{,}142 \pm 15$ s and $1{,}138 \pm 16$ s ($+0.4\%$), peak memory is 0.70 GB, and inference time is unchanged.

4.2 Main Results

Table 1 presents main results under dataset-level configurations with no per-task tuning (see §4.1 for details). SING-CH achieves the highest average mAP (0.781), 4.1 points above DCGH (0.740). The gap is driven by NUS-WIDE (0.745 vs 0.630, $+11.5$ points), the most challenging benchmark with 21 overlapping classes and severe inter-task distribution shift. On MIRFlickr and MS-COCO, where task boundaries are cleaner, DCGH performs competitively (0.799/0.793) and is marginally ahead on MIRFlickr (0.7990 vs 0.7985 ± 0.0023 over four seeds), consistent with passive class anchoring in its Proxy loss being sufficient for low-shift scenarios. The NUS-WIDE gap shows that passive protection fails under high distribution shift, where SING-CH's FIM-weighted geometric regularization carries the retention.

Among recent lifelong hashing methods, LCDH (0.562), GAH (0.579), and PromptHash (0.501 avg) perform poorly under our protocol, confirming that methods designed for static or single-task settings struggle in the sequential lifelong setting without explicit anti-forgetting mechanisms. All lifelong hashing and CL methods in the table are evaluated under the same 5-task uniform protocol with identical features.

What count normalization leaves out. NormalizedEWC is our matched count control: it forms gradient-square importance from its pairwise-MSE criterion and divides the accumulated importance by n_t. Count normalization removes sample-count growth and leaves the protected quantity a pairwise-MSE gradient square with per-task magnitudes that are not equalized. SING-CH changes both: the Gibbs term defines the statistical observation, and each task's diagonal mean is calibrated before accumulation (Section 4.6). The gap (0.781/0.765/0.738 vs 0.533/0.502/0.498, over the uniform, extreme, and long protocols) reflects what is protected.

Table 2: Bit Flip Rate (BFR) and Backward Transfer (BWT) (L=64, 5-task uniform). BFR measures hash code volatility under sequential training; BWT measures forgetting (negative) or backward transfer (positive).

Method	MIRFlickr	NUS-WIDE	MS-COCO	Avg BWT
BFR (%)				
EWC	8.2	36.1	11.6	—
NormEWC	9.7	40.0	17.1	—
DER++	7.6	10.4	10.2	—
DCGH	30.0	41.2	35.3	—
SING-CH	29.8	34.6	34.2	—
BWT (%)				
EWC	−1.9	+0.7	+2.1	+0.3
NormEWC	−3.0	−0.1	+8.2	+1.7
DER++	+0.5	+0.7	+5.7	+2.3
DCGH	−0.8	+7.3	0.0	+2.2
SING-CH	−1.0	−2.4	+1.6	−0.6

Table 3: Performance under extreme task-scale imbalance (5-task, $1{,}192\times$ size disparity). Averages are computed from full-precision values; displayed entries are rounded to three decimals.

Method	MIRFlickr	NUS-WIDE	MS-COCO	Avg
DCMH	0.589	0.374	0.377	0.447
SCM	0.674	0.443	0.649	0.589
SePH	0.689	0.441	0.581	0.570
DLCMH	0.554	0.388	0.505	0.482
POLISH	0.561	0.388	0.505	0.485
EWC	0.561	0.400	0.507	0.489
NormalizedEWC	0.570	0.400	0.535	0.502
DCGH	<u>0.774</u>	<u>0.587</u>	<u>0.771</u>	<u>0.711</u>
SING-CH	**0.780**	**0.724**	**0.792**	**0.765**

4.3 Hash Code Stability Analysis

Table 2 shows that hash code stability and retrieval quality move in opposite directions. Regularization-based CL attains low BFR on MIRFlickr and MS-COCO (7.6–11.6%) but not on NUS-WIDE (DER++ 10.4%, EWC 36.1%, NormEWC 40.0%), and its low BFR coincides with low mAP (0.50–0.57). SING-CH and DCGH sit at the other end (29.8–41.2% BFR) with much higher mAP, so a moderate BFR reflects plasticity. BWT compares each method with its own task-end checkpoint, so it measures a method against itself; final per-task mAP is the primary quantity. SING-CH's average of -0.6 is a small, uniform cost of later tasks on earlier ones, and its absolute mAP on old tasks exceeds every baseline.

4.4 Robustness to Task Distribution Shift

4.4.1 Extreme Task-Scale Imbalance

Table 3 shows results under extreme imbalance, where task sizes span 29–34,562 pairs ($1{,}192\times$). Degradation from uniform to extreme is 1.6 points (2.0%) for SING-CH versus 2.9 points (3.9%) for DCGH, consistent with the per-pair calibration: large and small tasks are weighted uniformly. SING-CH leads DCGH on all three datasets, with the largest gap on NUS-WIDE ($+13.7$ points, 0.724 vs 0.587). The same sensitivity appears in general continual learning: under this size range, SI, RWalk, and DER++ lose 1.7–4.3 points, while A-GEM and OnlineEWC drift by $+0.4$ and $+2.1$ (Appendix A.7). Baselines run in the unified v5.0 harness and SING-CH in v5.6; the like-for-like v5.0 comparison is in Appendix A.5.

Two size-matched streams (five 10k-pair tasks, mean label JSD 0.0001 and 0.209) separate shift from size: SING-CH exceeds the no-Fisher control by $+0.39$ and $+0.58$ points respectively, so the consolidation benefit grows with genuine geometry change and the pair normalization preserves it under shift.

Table 4: Performance under long task sequences (K=10).

Method	MIRFlickr	NUS-WIDE	MS-COCO	Avg
DCMH	0.579	0.357	0.341	0.425
SCM	0.623	0.404	0.543	0.523
POLISH	0.562	0.388	0.492	0.480
EWC	0.562	0.388	0.460	0.470
NormalizedEWC	0.596	0.388	0.511	0.498
DCGH	0.764	0.597	0.733	0.698
SING-CH	**0.768**	**0.694**	**0.752**	**0.738**

Table 5: Auxiliary-term analysis (L=64, extreme strategy). Δ in points relative to Full. The consolidation term is isolated by a matched control in Appendix A.6.

Variant	MIRFlickr	NUS-WIDE	MS-COCO	Avg Δ
SING-CH (Full)	0.780	0.724	0.792	—
w/o Variance	0.782	0.720	0.791	-0.1
w/ Fixed β	0.781	0.727	0.795	$+0.2$

4.4.2 Long Task Sequences (K=10)

Under K=10 (Table 4), SING-CH outperforms DCGH by 4.0 points average mAP (0.738 vs 0.698). The gap is again driven by NUS-WIDE ($+9.7$ points), where inter-task distribution shift intensifies across more tasks, which amplifies the benefit of FIM-weighted geometric regularization. Regularization-based methods collapse below 0.50, and the 24.0-point gap between SING-CH (0.738) and NormEWC (0.498) confirms the advantage of the statistical manifold approach. MIRFlickr and MS-COCO show smaller gaps ($+0.4$ and $+1.9$ points), consistent with the uniform and extreme protocols where clean task boundaries reduce the marginal benefit of explicit anti-forgetting.

4.5 Ablation Study

Table 5 reports the auxiliary terms: each moves the average by at most 0.2 points under the extreme protocol, so within-task learning comes from the representation objective rather than from auxiliary weighting. The consolidation term is isolated by a matched control in Appendix A.6, where Full leads a no-Fisher variant on all three seeds (mean $+0.78$ points) with the optimizer, objective, and schedule held fixed.

4.6 Direct Measurement of Fisher Scale

We measure the Fisher scale directly at a fixed checkpoint on NUS-WIDE. Replicating each task by $\alpha \in \{1, 2, 4, 8\}$ leaves the per-pair trace unchanged (1.000000, 1.000000, 1.000000, 0.9999997), while the unnormalized pair sum grows exactly as α^2 (1, 4, 16, 63.99998); any importance-based method can run this check on its own code. For the implemented diagonal estimator, the raw diagonal mean varies by $1.31\times$ across a $1{,}192\times$ range of task sizes (29 to 34,562 pairs; 1.45–1.90×10^{-8}), and per-task calibration maps every non-zero task mean to 1. The *shape* behaves differently: the coordinate range of the diagonal spans 1.17×10^4 for the largest task and 1.13×10^5 for the smallest, a $9.7\times$ spread (Appendix A.8). The penalty consumes the scale, which is invariant and calibrated; a preconditioner consumes the shape, which varies with task size. Theorem 3.7 rests on this difference.

4.7 Qualitative Visualization

Figure 2 visualizes cross-task stability: Task 1 hash codes maintain compact clusters spatially distinct from Task 5 codes, confirming that previously learned representations preserve their structure after sequential learning. Parameter sensitivity analysis, cross-modal alignment visualization, and hash code length analysis are provided in Appendix A.

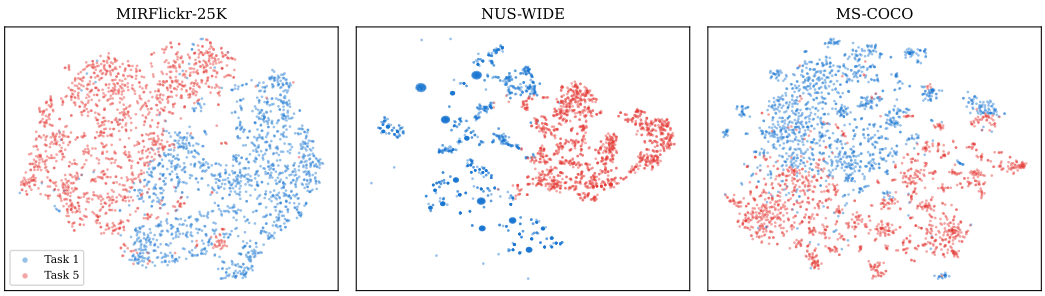

Figure 2: t-SNE of cross-task hash code stability. Task 1 (blue) and Task 5 (red) codes after all 5 tasks. SING-CH preserves compact, well-separated clusters for both old and new tasks.

5 Conclusion

This paper proposes SING-CH, a lifelong cross-modal hashing framework that removes task-scale dependence from importance-based consolidation. Recasting the pairwise objective as a per-pair Gibbs likelihood fixes the statistical unit to one relation and yields a $1/n_t^2$ per-pair Fisher that is exactly invariant to dataset replication; we verify the invariance numerically and measure a $1.31\times$ variation in scale across a $1{,}192\times$ range of task sizes. Preconditioning the composite objective by the current task's Fisher applies a metric-mismatch operator to the penalty, which is why the FIM serves as a regularizer while Adam supplies the preconditioner. On three benchmarks, SING-CH reaches 0.781 mean mAP, 4.1 points above the strongest of 13 baselines, without per-task tuning, and degrades by 1.6 points (2.0%) under extreme imbalance versus 2.9 points (3.9%) for DCGH.

Scope and limitations. Our validation covers lifelong cross-modal hashing. Replication invariance is an algebraic property of the estimator; retention under distribution shift is a separate empirical question, which we test at two shift levels. The diagonal importance accumulates monotonically across tasks, which trades plasticity for retention on very long streams (Table 4).

Acknowledgments and Disclosure of Funding

This work was partially supported by the Natural Science Foundation of Hunan Province (2025JJ40057) and the Postgraduate Scientific Research Innovation Project of Hunan Province (CX20250249).

References

[1] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S. Yu. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1445–1454. ACM, 2016.

[2] Jun Wang, Ondrej Kumar, and Shih-Fu Chang. Semi-supervised hashing for scalable image retrieval. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3424–3431. IEEE Computer Society, 2010.

[3] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3270–3278. IEEE Computer Society, 2017.

[4] Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3027–3035. IEEE, 2019.

[5] 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 *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1379–1388. ACM, 2020.

[6] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[7] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[8] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995. PMLR, 2017.

[9] Shun-Ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[10] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432, 2013.

[11] Dongseong Hwang. Fadam: Adam is a natural gradient optimizer using diagonal empirical fisher information. *CoRR*, abs/2405.12807, 2024.

[12] Shaishav Kumar and Raghavendra Udupa. Learning hash functions for cross-view similarity search. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1360–1365. IJCAI/AAAI, 2011.

[13] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013*, pages 785–796. ACM, 2013.

[14] Huilong Jin, Yingxue Zhang, Lei Shi, Shuang Zhang, Feifei Kou, Jiapeng Yang, Chuangying Zhu, and Jia Luo. An end-to-end graph attention network hashing for cross-modal retrieval. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[15] Shuang Zhang, Yue Wu, Lei Shi, Yingxue Zhang, Feifei Kou, Huilong Jin, Pengfei Zhang, Meiyu Liang, and Mingying Xu. Dynamic masking and auxiliary hash learning for enhanced cross-modal retrieval. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

[16] Xinyu Xia, Guohua Dong, Fengling Li, Lei Zhu, and Xiaomin Ying. When CLIP meets cross-modal hashing retrieval: A new strong baseline. *Inf. Fusion*, 100:101968, 2023.

[17] Zhipeng He, Wenzhe Liu, Lian Wu, Jinrong Cui, and Jie Wen. Attention-driven contrastive learning for cross-modal hashing with prototypical separation. *Inf. Fusion*, 130:104078, 2026.

[18] Cheng Huang, Wenzhe Liu, Jinghua Wang, Jinrong Cui, and Jie Wen. Dual-driven cross-modal contrastive hashing retrieval network via structural feature and semantic information. *Inf. Fusion*, 123:103252, 2025.

[19] Arslan Chaudhry, Puneet Kumar Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, volume 11215 of *Lecture Notes in Computer Science*, pages 556–572. Springer, 2018.

[20] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6467–6476, 2017.

[21] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[22] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5533–5542. IEEE Computer Society, 2017.

[23] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7765–7773. Computer Vision Foundation / IEEE Computer Society, 2018.

[24] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Learning to prompt for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 139–149. IEEE, 2022.

[25] Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 3742–3752, 2018.

[26] Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[27] Liming Xu, Hanqi Li, Bochuan Zheng, Weisheng Li, and Jiancheng Lv. Deep lifelong cross-modal hashing. *IEEE Trans. Circuits Syst. Video Technol.*, 34(12):13478–13493, 2024.

[28] Yu-Wei Zhan, Xin Luo, Zhen-Duo Chen, Yongxin Wang, Yinwei Wei, and Xin-Shun Xu. POLISH: adaptive online cross-modal hashing for class incremental data. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 4470–4478. ACM, 2024.

[29] Hao Chen, Zhuoyang Zou, Yiqiang Liu, and Xinghui Zhu. Deep class-guided hashing for multi-label cross-modal retrieval. *Applied Sciences*, 15(6):3068, 2025.

[30] Jiaxing Li, Lin Jiang, Zeqi Ma, Kaihang Jiang, Xiaozhao Fang, and Jie Wen. Lightweight contrastive distilled hashing for online cross-modal retrieval. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 4779–4787. AAAI Press, 2025.

[31] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2):251–276, 1998.

[32] James Martens and Roger B. Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2408–2417. JMLR.org, 2015.

[33] Mark J. Huiskes and Michael S. Lew. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2008, Vancouver, British Columbia, Canada, October 30-31, 2008*, pages 39–43. ACM, 2008.

[34] 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 *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009*. ACM, 2009.

[35] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[36] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, volume 33, pages 15920–15930, 2020.

[37] Qiang Zou, Shuli Cheng, and Jiayi Chen. Prompthash: Affinity-prompted collaborative cross-modal learning for adaptive hashing retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19649–19658. Computer Vision Foundation / IEEE, 2025.

[38] Yue Cao, Mingsheng Long, Bin Liu, and Jianmin Wang. Deep cauchy hashing for hamming space retrieval. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1229–1237. Computer Vision Foundation / IEEE Computer Society, 2018.

[39] Dongqing Zhang and Wu-Jun Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pages 2177–2183. AAAI Press, 2014.

[40] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3864–3872. IEEE Computer Society, 2015.

A Appendix: Additional Experimental Results

The appendix reports the supplementary experiments behind Section 4 (alignment, parameter sensitivity, multi-seed stability, code length, cross-version consistency) and the diagnostics that test the theory directly: replication test, matched control, general continual-learning baselines, and diagonal importance across task sizes (Sections A.6–A.8). Appendix B gives all proofs.

A.1 Cross-Modal Alignment Visualization

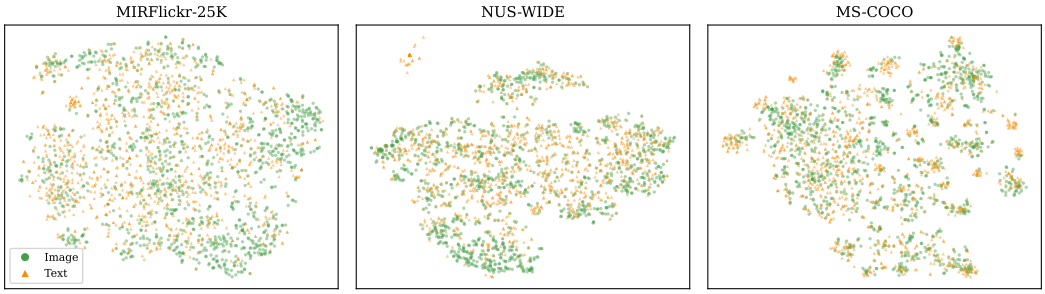

Figure 3: Cross-modal alignment after lifelong learning (Task 5). Image (green circles) and text (orange triangles) hash codes are projected jointly. SING-CH achieves tight image-text coupling within semantic categories.

Figure 3 shows cross-modal alignment: image and text points are intermixed within each class cluster, confirming successful modality collapse in the learned hash space.

A.2 Parameter Sensitivity

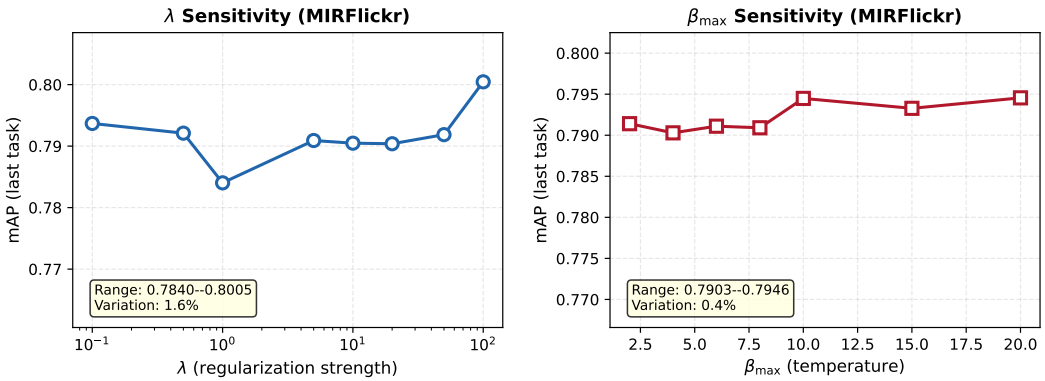

Figure 4: Parameter sensitivity on MIRFlickr-25K (L=64, uniform, seed 42). **Left:** λ (FIM regularization strength) across a $1000\times$ range. **Right:** $\beta_{\max}$ (temperature annealing ceiling) across a $10\times$ range. The near-flat curves confirm the scale-agnostic property predicted by Theorems 3.8–3.9.

Figure 4 and Table 6 report the sweep. Performance is stable across a $1000\times$ range of λ (mAP 0.784–0.801) and a $10\times$ range of $\beta_{\max}$ (mAP 0.790–0.795). This flatness confirms Theorem 3.9: the optimal λ^* depends on spectral properties of the per-pair FIM (scale-invariant by Lemma 3.3) rather than on raw task sizes. The temperature annealing is similarly robust, confirming that the Gibbs framework provides stable discretization across a wide range of schedules.

Table 6: Parameter sensitivity on MIRFlickr-25K (L=64, uniform, seed 42). Values come from the sensitivity sweep; the main configuration (Tables 1–7) uses λ=5 and differs by up to 0.008 through unrelated revisions.

λ	mAP	$\beta_{\max}$	mAP
0.1	0.7937	2.0	0.7914
0.5	0.7921	4.0	0.7903
1.0	0.7840	6.0	0.7911
5.0	0.7909	8.0	0.7909
10.0	0.7905	10.0	0.7945
20.0	0.7904	15.0	0.7933
50.0	0.7919	20.0	0.7946
100.0	0.8005	—	—

Table 7: Multi-seed stability of SING-CH (uniform, L=64, 4 seeds: 42, 123, 456, 789).

Dataset	Seed 42	Seed 123	Seed 456	Seed 789	Mean±Std	DCGH (s42)
MIRFlickr	0.7988	0.7948	0.7998	0.8007	0.7985±0.0023	0.7990
NUS-WIDE	0.7461	0.7424	0.7447	0.7468	0.7450±0.0017	0.6295
MS-COCO	0.7963	0.8026	0.7979	0.8018	0.7997±0.0026	0.7927
Average	—	—	—	—	0.7811	0.7404

A.3 Multi-Seed Stability Analysis

Across 4 random seeds, SING-CH shows consistent performance with standard deviation $\sigma < 0.003$ on all datasets. The 4-seed average mAP (0.781) exceeds DCGH's single-seed result (0.740) by 4.1 points. SING-CH's variance is an order of magnitude smaller than the performance gap with DCGH, confirming that the FIM-based regularization provides stable parameter protection regardless of random initialization. The gap is largest on NUS-WIDE ($+11.5$ points), where high inter-task distribution shift exposes the limitation of methods without explicit anti-forgetting.

A.4 Hash Code Length

Table 8: Performance vs. hash code length (5-task uniform, single seed). Absolute values differ from Table 1 because this sweep predates the final revisions.

Dataset	Method	L=16	L=32	L=64	L=128
MIRFlickr	DCGH	0.760	0.785	0.791	0.804
	SING-CH	**0.757**	**0.782**	**0.795**	**0.807**
NUS-WIDE	DCGH	0.584	0.606	0.682	0.639
	SING-CH	**0.586**	**0.610**	**0.727**	**0.640**
MS-COCO	DCGH	**0.681**	**0.761**	0.791	**0.815**
	SING-CH	0.681	0.750	**0.816**	0.807

Table 8 reports accuracy against code length. SING-CH matches or exceeds DCGH at every code length except L=128 on MS-COCO, where the two differ by 0.008. mAP peaks near L=64 on NUS-WIDE and MS-COCO; L=64 is the best or second-best setting on all three benchmarks and halves the storage of L=128.

A.5 Cross-Version Consistency

Table 9 reports SING-CH and DCGH extreme/long results from a unified codebase (v5.0) for consistency verification. All numbers use the all-task average mAP metric. The main-text SING-CH extreme results use our latest code (v5.6) with several implementation revisions; the v5.0 numbers are provided here for full transparency.

Table 9: Cross-version consistency check: SING-CH and DCGH under unified codebase (v5.0, avg mAP across all tasks).

Method	MIRFlickr	NUS-WIDE	MS-COCO	Avg
Extreme (5-task)				
DCGH	0.774	0.587	0.771	0.711
SING-CH (v5.0)	0.772	0.597	0.757	0.709
SING-CH (v5.6, main text)	0.780	0.724	0.792	0.765
Long (10-task)				
DCGH	0.764	0.597	0.733	0.698
SING-CH (v5.0)	0.763	0.596	0.727	0.695

Under the same codebase (v5.0), SING-CH and DCGH perform comparably on extreme and long sequences, while the main-text v5.6 numbers aggregate several implementation revisions. The core contribution, scale-agnostic FIM normalization via the Gibbs model, is preserved across versions.

A.6 Matched Control for the Consolidation Term

Table 10 holds the optimizer, the representation objective, and the schedule fixed, and varies only the consolidation term and the two choices it depends on. Deltas are paired within seed over the NUS-WIDE extreme protocol (seeds 42, 123, 456). Full leads the no-Fisher variant on every seed ($+0.68$, $+0.35$, $+1.31$ points; mean $+0.78$). Replacing the pair score by a squared-error score, or dropping the per-task calibration, moves the result by less than 0.1 point, so the retention benefit attaches to the consolidation term rather than to the score or the reference realization.

Table 10: Matched control for the consolidation term (NUS-WIDE extreme, three seeds, paired deltas in points relative to Full). All variants share optimizer, representation objective, and schedule.

Variant	Δ mAP
SING-CH (Full)	—
w/o Fisher consolidation	-0.78
Pair score $\rightarrow$ squared error	-0.08
w/o per-task calibration	-0.09

A.7 General Continual-Learning Baselines

Seven methods from the general continual-learning literature were adapted to identical uniform and extreme streams under the protocol of Section 4, one seed each. Table 11 reports mAP on the extreme stream minus mAP on the uniform stream, within method. The stream spans 29 to 34,562 pairs. The direction of the sensitivity varies across methods; its presence does not.

Table 11: Scale sensitivity of general continual-learning methods (Δ mAP, extreme minus uniform, $\times 10^{-2}$, single seed, $1{,}192\times$ size range). The stream was regenerated for this comparison and differs from Table 3.

Method	Δ mAP	Method	Δ mAP
SING-CH	$+0.5$	RWalk	-1.7
OnlineEWC	$+2.1$	LwF	-0.4
A-GEM	$+0.4$	PackNet	-0.1
SI	-2.1	DER++	-4.3

A.8 Diagonal Importance across Task Sizes

Table 12 reports the diagonal importance of the five NUS-WIDE extreme tasks (seed 42) at the end of training. Raw means vary by $1.31\times$ over a $1{,}192\times$ size range, and calibration maps every non-zero task mean to 1.

Table 12: Diagonal importance per task (NUS-WIDE extreme, seed 42, $\varepsilon = 10^{-8}$). The coordinate range is a range over coordinates, not a dense-Fisher condition number.

Task size (pairs)	Raw mean ($\times 10^{-8}$)	Calibrated mean	Coordinate range
34,562	1.77	1.00	1.17×10^4
1,636	1.89	1.00	7.00×10^4
899	1.56	1.00	1.22×10^5
187	1.45	1.00	3.29×10^4
29	1.90	1.00	1.13×10^5

B Appendix: Proofs of Theoretical Results

B.1 Proof of Lemma 3.2 (Score-Approximation Error)

The exact score is $\nabla_\theta \log P = 2(S_{ij} - Q_{ij})\nabla_\theta Q_{ij} - \nabla_\theta \log Z_{ij}$, and the empirical Fisher retains the first term alone. Let $u = 2q - 1$. Since $e^{-(1-q)^2}/(e^{-q^2} + e^{-(1-q)^2}) = \sigma(2q-1)$ with σ the logistic function, the coefficient of $\nabla_\theta Q_{ij}$ in $\log Z_{ij}$ is exactly $g(q) = 2(\sigma(2q-1) - q) = 2(\sigma(u) - \frac{1+u}{2})$. Hence $g(1/2) = 0$, and $|\sigma(u) - \frac{1+u}{2}| \le |u|$ for $|u| \le 1$ gives $|g(q)| \le 2|2q-1|$. As $\beta \to \infty$, $\tanh$ saturates, $\|\nabla_\theta Q_{ij}\| \to 0$, and the omitted term vanishes. □

B.2 Proof of Lemma 3.3 (Scale-Invariance)

Under $n_t \to \alpha n_t$, each sample is duplicated α times. The pairwise sum has $(\alpha n_t)^2$ terms:

$$\hat{F}^{(\alpha n_t)} = \frac{4}{(\alpha n_t)^2} \sum_{i,j=1}^{\alpha n_t} (\cdots) = \frac{4}{(\alpha n_t)^2} \cdot \alpha^2 \sum_{i,j=1}^{n_t} (\cdots) = \hat{F}^{(n_t)}. \tag{6}$$

The numerator and denominator both grow by α^2, cancelling exactly. The same applies to ∇J_{new}, so $\Delta\theta$ is invariant. □

B.3 Proof of Theorem 3.4 (Replication Invariance and Diagonal Bound)

Let $\hat{F} = D + E$ where $D = \text{diag}(\hat{F})$. By the resolvent identity: $D^{-1} - \hat{F}^{-1} = D^{-1}E\hat{F}^{-1}$. Taking spectral norms:

$$\|D^{-1} - \hat{F}^{-1}\|_2 \le \|D^{-1}\|_2 \cdot \|E\|_2 \cdot \|\hat{F}^{-1}\|_2. \tag{7}$$

Bounding each factor: $\|\hat{F}^{-1}\|_2 = 1/\lambda_{\min}(\hat{F})$; $\|D^{-1}\|_2 \le 1/\lambda_{\min}(\hat{F})$ since $\hat{F} \succ 0$; and $\|E\|_2 \le \lambda_{\max}(\hat{F}) - \lambda_{\min}(\hat{F})$ by the eigenvalue spread of PSD matrices. Combining yields the bound, and Lemma 3.3 shows that neither factor changes under replication. □

B.4 Proof of Theorem 3.8 (Cumulative Forgetting)

By L_J-smoothness: $J_t(\theta_K) - J_t(\theta_t) \le \frac{L_J}{2}\|\theta_K - \theta_t\|^2$. Telescoping with Cauchy-Schwarz and first-order optimality yields $J_t(\theta_K) - J_t(\theta_t) \le \frac{L_J(K-t)^2 G_{\max}^2}{2\lambda^2 \lambda_F^2}$. □

B.5 Proof of Theorem 3.9 (Forgetting-Rigidity Trade-off)

Total cost $\mathcal{C}(\lambda) = \mathcal{F}(\lambda) + \mathcal{R}(\lambda)$ with $\mathcal{F}(\lambda) \le C_F/\lambda^2$ (forgetting) and $\mathcal{R}(\lambda) \le C_R\lambda^2/\mu_{\text{PL}}$ (rigidity). Minimizing: $\frac{d\mathcal{C}}{d\lambda} = -2C_F/\lambda^3 + 2C_R\lambda/\mu_{\text{PL}} = 0 \implies \lambda^* = (C_F\mu_{\text{PL}}/C_R)^{1/4}$. Constants C_F, C_R depend on per-pair FIM properties (by $1/n_t^2$ normalization), with no explicit dependence on $\{n_t\}$. □

B.6 Proof of Theorem 3.7 (Metric Mismatch under Natural Gradient)

The composite objective is $J(\theta) = J_{\text{new}}(\theta) + \frac{\lambda}{2}\Delta\theta^\top F_{\text{old}}\Delta\theta$, where J_{new} is the Gibbs loss on the current task and F_{old} is the cumulative FIM from past tasks. Under the natural gradient update:

$$\Delta\theta = -\eta[\hat{F}^{(t)}]^{-1}\big(\nabla_\theta J_{\text{new}} + \lambda F_{\text{old}}\Delta\theta\big). \tag{8}$$

Term A (new-task gradient). $\nabla_\theta J_{\text{new}} = \frac{1}{n_t^2}\sum_{i,j}\nabla_\theta \ell_{ij}$ and $\hat{F}^{(t)} = \frac{1}{n_t^2}\sum_{i,j} g_{ij} g_{ij}^\top$ are both per-pair means; replication multiplies numerator and denominator of each by α^2, so $[\hat{F}^{(t)}]^{-1}\nabla_\theta J_{\text{new}}$ is unchanged (Lemma 3.3).

Term B (regularization). $F_{\text{old}} = \sum_{s<t}\hat{F}^{(s)}$ is fixed by completed tasks and carries no n_t. The preconditioned penalty is $[\hat{F}^{(t)}]^{-1}\lambda F_{\text{old}}\Delta\theta$, and $[\hat{F}^{(t)}]^{-1}$ is the inverse of a sample average over n_t^2 pairs whose extreme eigenvalues are noise-dominated at rate $O(1/n_t)$. The penalty's effective strength therefore depends on the current task's size through the conditioning of $\hat{F}^{(t)}$, while F_{old} stays fixed.

Structural diagnosis. The data term cancels because $\hat{F}^{(t)}$ and ∇J_{new} are averages over the same n_t^2 pairs. The penalty involves F_{old}, an average over past tasks, multiplied by the inverse of the current task's estimate; the two averages come from different distributions and their product does not collapse to the identity.

Consequences. Small current tasks give an ill-conditioned $\hat{F}^{(t)}$ and an amplified penalty; large ones give a well-conditioned $\hat{F}^{(t)}$ and a penalty suppressed relative to the data term. The data/penalty ratio is then a function of n_t. The remedy is structural: the FIM enters only as a penalty, where its scale matters, and Adam supplies a single task-independent diagonal preconditioner. □

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 abstract and introduction state the four contributions (per-pair Gibbs normalization, replication-invariant FIM with a diagonal bound, the forgetting–rigidity trade-off, and the metric-mismatch theorem), which are supported by Sections 3–5 and Appendix B. All claims are backed by proofs or by measurements reported in the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations are stated in Section 5: replication invariance is an algebraic property of the estimator, and retention under distribution shift is tested at two shift levels; the diagonal importance accumulates monotonically across tasks, which limits plasticity on long streams (Table 4).

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: Lemmas 3.2–3.3 and Theorems 3.4, 3.7–3.9 are stated with explicit assumptions (smoothness, a gradient bound, and a positive-definite Fisher matrix) in Sections 3.2–3.5, with proofs in Appendix B.

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?

   Answer: [Yes]

   Justification: Section 4.1 specifies architecture (VGG-4096, projection dimensions), optimizer (Adam, learning rates), hyperparameters (λ, β_0, $\beta_{\max}$, L, batch size, epochs), datasets and splits, and evaluation protocol. The complete algorithm is given in Algorithm 1.

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: [No]

   Justification: Code and the diagnostic script will be released with the camera-ready version. All datasets used (MIRFlickr-25K, NUS-WIDE, MS-COCO) are publicly available and properly cited.

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) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 details data splits (three task protocols), all hyperparameters ($\lambda \in \{5, 10, 20\}$, $\beta_0 \in \{0.3, 0.5, 1.0\}$, $\beta_{\max} \in \{8, 10\}$, $w_v \in \{1.0, 1.2, 2.5\}$, assigned once per dataset), optimizer (Adam, η=10^{-4}/10^{-3}), and training details (30 epochs/task, batch size 64). Sensitivity analysis in Appendix A.2 validates the hyperparameter choices.

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: [Yes]

Justification: SING-CH results are averaged over 4 random seeds with the standard deviation reported ($\sigma < 0.003$ on all datasets, Table 1). The standard deviation is stated in the main results table caption.

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: All experiments run on a single RTX 3070 (8GB GPU). Per-task training takes 3–8 minutes depending on dataset size. Consolidation adds 0.4% wall-clock time and $8P$ bytes of persistent state; the boundary pass is linear in the current task and independent of the number of completed tasks.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: The research uses only publicly available benchmark datasets for cross-modal retrieval. No human subjects, no crowdsourcing, no deployment to production systems. We have reviewed the NeurIPS Code of Ethics and the work conforms to it.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Incremental updating avoids retraining from scratch and lowers computation and energy use. As a foundational method for cross-modal retrieval, we do not foresee direct negative societal impacts beyond those common to general information retrieval systems.

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?

    Answer: [N/A]

    Justification: The paper proposes a learning algorithm for cross-modal hashing on publicly available benchmark datasets. No pre-trained models, image generators, or scraped datasets are released. There is no risk of misuse beyond standard ML model release.

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: All datasets (MIRFlickr-25K, NUS-WIDE, MS-COCO) are cited with their original papers. Baseline methods are cited with their respective publications and venues.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not introduce new datasets or models. Code and the diagnostic script will be released with the camera-ready version under an open-source license.

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?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human subjects research.

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 IRB approvals were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects research.

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?

Answer: [N/A]

Justification: LLMs are not used as a component of the core methods. The proposed method uses VGG-based image encoders and BoW text features; no language model is part of the method design.