UNIFIED SEMANTIC PRODUCT QUANTIZATION WITH CATEGORY-AWARE IMPLICIT CODEBOOKS FOR CROSS-MODAL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal retrieval (CMR) requires compact and discriminative representations for efficient large-scale search. Although product quantization (PQ) is effective for compact representation learning, existing PQ-based CMR methods typically treat codebooks primarily as feature-approximation prototypes, limiting the direct incorporation of semantic supervision into codebook learning. To address this issue, we propose **P2PQ**, a unified semantic PQ framework with category-aware implicit codebooks for CMR. Instead of independently parameterizing quantization codebooks and semantic predictors, P2PQ shares a single matrix between these two roles, so that the codebook cardinality is induced by the dimensionality of the semantic prediction space rather than specified independently. This dual-role parameterization allows category-level supervision to directly shape the quantization prototypes without imposing label-constrained codeword assignments, thereby intrinsically coupling discrete quantization with semantic discrimination. Furthermore, P2PQ learns a shared latent semantic space for heterogeneous modality alignment and preserves geometric consistency between continuous and quantized representations through an implicit label-guided affinity formulation without explicit graph construction. Extensive experiments on three benchmark datasets demonstrate that P2PQ consistently outperforms representative shallow CMR methods under various code lengths, validating its effectiveness and scalability.

1 INTRODUCTION

Cross-modal retrieval (CMR) Wang et al. (2025); Pu et al. (2025); Chen et al. (2026) aims to retrieve semantically relevant instances across heterogeneous modalities, such as images and texts. In large-scale retrieval scenarios, directly storing and comparing high-dimensional continuous representations usually incurs considerable memory and computational costs. To improve retrieval efficiency, compact representation learning has attracted increasing attention, among which hashing and product quantization (PQ) are two representative solutions Jegou et al. (2010); Hu et al. (2024); Sun et al. (2026); Su et al. (2026). Compared with binary hashing, PQ represents data by compositional codewords and supports efficient similarity search with storage-accuracy trade-offs Jegou et al. (2010); Chen et al. (2022); Huijben et al. (2024). However, although PQ has been widely studied in approximate nearest neighbor search, its application to CMR remains relatively underexplored.

Existing PQ-based CMR methods mainly learn codebooks to approximate continuous features, where the codewords primarily serve as quantization prototypes Long et al. (2016); Wang et al. (2022). Such a design is effective for reducing quantization error, but it is not sufficiently tailored to CMR, where compact codes are expected to be not only storage-efficient but also semantically discriminative Zhou et al. (2020). Moreover, conventional PQ methods often adopt a fixed-size codebook in each subspace, whose codewords are optimized primarily for feature approximation and are not directly shaped by category-level semantic supervision. As a result, the learned codebooks may contain redundant or semantically ambiguous codewords, while semantic supervision is imposed only indirectly on the resulting representations rather than being explicitly incorporated into codebook learning.

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.

Motivated by this observation, we rethink the role of codebooks in PQ-based CMR. Instead of treating codebooks merely as feature approximation prototypes, we propose category-aware implicit codebooks whose size is tied to the number of semantic categories. More importantly, the same codebook matrix is shared between subspace quantization and label regression, allowing category-level supervision to directly participate in codebook optimization. Therefore, category awareness here refers to the semantic guidance imposed on codebook learning, rather than a hard constraint that assigns each sample to the codeword indexed by one of its labels. This design directly couples discrete quantization with semantic discrimination and yields more discriminative quantized representations.

Based on this idea, we propose P2PQ, a unified PQ framework with implicit semantic codebooks for CMR. Specifically, P2PQ first learns a shared latent semantic space, where modality-specific projections, orthogonal transformations, shared semantic bases, and modality-adaptive weights are jointly optimized to alleviate heterogeneous modality discrepancies. This design is related to semantic alignment in cross-modal representation learning Tang et al. (2016); Chen et al. (2020), but further incorporates modality-adaptive semantic modulation within a unified PQ-oriented objective. Then, a joint subspace quantization and label regression mechanism is introduced, in which the category-aware codebooks are shared by discrete reconstruction and semantic regression, thereby bridging quantization learning and discriminative semantic prediction. Furthermore, to preserve semantic relationships during quantization, we develop an implicit label-guided affinity formulation that aligns the pairwise relationships induced by continuous representations and their quantized counterparts without explicitly constructing an affinity graph Dong et al. (2022); Wang et al. (2021); Zheng et al. (2026). All these components are integrated into a unified objective and solved by an efficient alternating optimization algorithm.

The main contributions of this work are summarized as follows:

- We propose a category-aware implicit codebook learning mechanism for PQ-based CMR. The codebook size in each subspace is determined by the number of semantic categories, reducing the number of codewords while allowing codewords to serve as both quantization prototypes and semantic regression weights.
- We develop a unified semantic alignment and quantization framework that jointly learns modality-specific transformations, shared semantic structures, modality-adaptive weights, and compact quantized representations, enabling heterogeneous modalities to be aligned in a common semantic space.
- We introduce an implicit label-guided geometric preservation strategy to maintain pairwise relationship consistency between continuous and quantized representations without explicit affinity graph construction.
- Extensive experiments on three benchmark datasets show that P2PQ consistently outperforms state-of-the-art shallow CMR methods under various code lengths.

2 METHODOLOGY

2.1 NOTATIONS

Before formulating the optimization problem, we first introduce the main notations. Bold uppercase letters denote matrices, bold lowercase letters denote vectors, and non-bold letters denote scalars. Let P be the number of modalities, N the number of instances, and c the number of semantic categories. In the proposed category-aware codebook design, c also determines the number of codewords in each quantization subspace. For the p-th modality, $\mathbf{X}_p \in \mathbb{R}^{d_p \times N}$ denotes the nonlinear feature matrix obtained by RBF kernel mapping Sun et al. (2024a); Liu et al. (2025), where d_p is the mapped feature dimension. The shared latent dimension is r, which is divided into M subspaces with equal dimension $d = r/M$. The Frobenius norm is denoted by $\|\cdot\|_{\mathrm{F}}$, $\mathbf{1}_N$ is an all-one vector, $\mathbf{I}$ is an identity matrix, and $\odot$ denotes the Hadamard product. Let $\mathbf{L} \in \{0,1\}^{N \times c}$ be the label matrix and $\bar{\mathbf{L}} \in \mathbb{R}^{N \times c}$ be its normalized version.

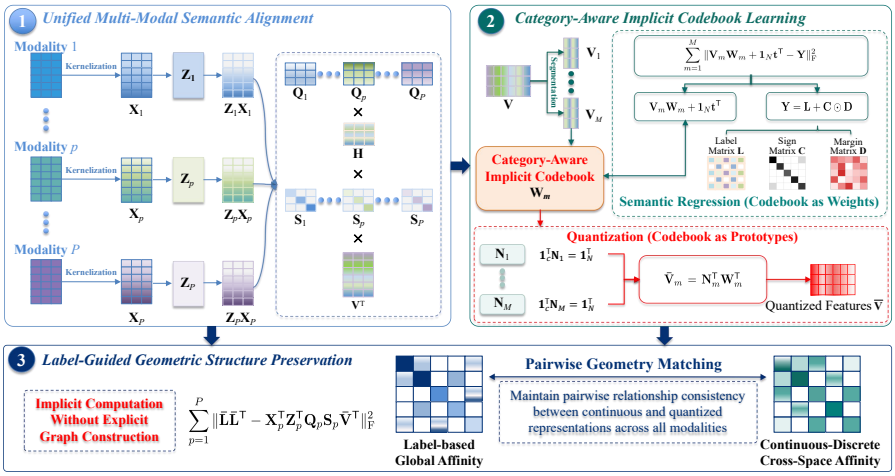


Figure 1: The framework of the proposed P2PQ.

2.2 Problem Formulation

As shown in Fig. 1, given multi-modal data $\{\mathbf{X}_p\}_{p=1}^{P}$ and label matrix $\mathbf{L}$, P2PQ jointly learns semantically aligned latent representations and compact quantized codes for efficient CMR. Unlike conventional PQ methods with predefined fixed-size codebooks, P2PQ constructs category-aware implicit codebooks whose size is determined by the number of semantic categories, enabling codewords to serve both discrete reconstruction and semantic regression.

Unified Multi-Modal Semantic Alignment. Different modalities usually lie in heterogeneous feature spaces with distinct distributions, making direct semantic comparison difficult. To alleviate such heterogeneity, we project each modality into a shared latent semantic space and further compensate for modality-dependent geometric variations through an orthogonal transformation. Specifically, for the p-th modality, we introduce a projection matrix $\mathbf{Z}_p \in \mathbb{R}^{r\times d_p}$ and an orthogonal transformation $\mathbf{Q}_p \in \mathbb{R}^{r\times r}$ with $\mathbf{Q}_p^{\mathsf{T}}\mathbf{Q}_p = \mathbf{I}_r$. The projected modality features are reconstructed by a shared semantic factorization $\mathbf{Q}_p\mathbf{H}\mathbf{S}_p\mathbf{V}^{\mathsf{T}}$, where $\mathbf{H} \in \mathbb{R}^{r\times r}$ denotes the shared semantic basis, $\mathbf{S}_p$ is a diagonal modality-adaptive weighting matrix, and $\mathbf{V} \in \mathbb{R}^{N\times r}$ is the common semantic representation. The semantic alignment objective is formulated as

$$\min_{\substack{\mathbf{Z}_p,\mathbf{Q}_p,\\ \mathbf{H},\mathbf{S}_p}} \sum_{p=1}^{P} \|\mathbf{Z}_p\mathbf{X}_p - \mathbf{Q}_p\mathbf{H}\mathbf{S}_p\mathbf{V}^{\mathsf{T}}\|_{\mathrm{F}}^2, \quad \text{s.t. } \mathbf{Q}_p^{\mathsf{T}}\mathbf{Q}_p = \mathbf{I}_r. \tag{1}$$

This formulation jointly learns modality-specific projections, orthogonal semantic transformations, shared semantic bases, and modality-adaptive weights. The learned representation $\mathbf{V}$ further serves as the bridge between semantic alignment and subsequent PQ, directly affecting the semantic consistency and discriminative capability of the quantized representation.

Category-Aware Implicit Codebook Learning. To enable efficient retrieval, we quantize the learned representation $\mathbf{V}$ into compact discrete codes. Different from conventional PQ methods that usually employ predefined fixed-size codebooks for feature approximation, we construct a category-aware implicit codebook in each subspace, whose size is determined by the number of semantic categories. Specifically, we decompose $\mathbf{V}$ into M subspaces, i.e., $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_M]$, where $\mathbf{V}_m \in \mathbb{R}^{N\times d}$. For the m-th subspace, $\mathbf{W}_m \in \mathbb{R}^{d\times c}$ denotes an implicit codebook containing c codewords, where its size is tied to the semantic label dimension. The semantic awareness of the codebook is introduced through its shared use in quantization and the subsequent label-regression objective. Let $\mathbf{N}_m \in \{0,1\}^{c\times N}$ be the assignment matrix. Each sample selects exactly one codeword in each subspace, enforced by

$$\mathbf{1}_c^{\mathsf{T}}\mathbf{N}_m = \mathbf{1}_N^{\mathsf{T}}. \tag{2}$$

The local quantized representation is $\bar{\mathbf{V}}_m = \mathbf{N}_m^{\mathsf{T}}\mathbf{W}_m^{\mathsf{T}} \in \mathbb{R}^{N\times d}$, and the global quantized representation is $\bar{\mathbf{V}} = [\bar{\mathbf{V}}_1, \bar{\mathbf{V}}_2, \ldots, \bar{\mathbf{V}}_M] \in \mathbb{R}^{N\times r}$. Although each subspace contains only c codewords, the compositional structure provides c^M possible codeword combinations, maintaining sufficient representation capacity.

To make the codebook semantically discriminative, we further let $\mathbf{W}_m$ participate in flexible label regression. The semantic regression objective is

$$\sum_{m=1}^{M} \|\mathbf{V}_m\mathbf{W}_m + \mathbf{1}_N\mathbf{t}^\top - \mathbf{Y}\|_{\mathrm{F}}^2, \tag{3}$$

where $\mathbf{t} \in \mathbb{R}^{c\times 1}$ is a learnable bias vector. $\mathbf{Y} = \mathbf{L} + \mathbf{C} \odot \mathbf{D}$ and $\mathbf{D} \geq 0$. Here, $\mathbf{C}$ is a sign matrix defined as

$$\mathbf{C}_{ij} = \begin{cases} +1, & \text{if } \mathbf{L}_{ij} = 1, \\ -1, & \text{if } \mathbf{L}_{ij} = 0. \end{cases} \tag{4}$$

Since the j-th column of $\mathbf{W}_m$ simultaneously acts as the j-th quantization prototype and the regression weight for the j-th category, category-level supervision directly shapes the codebook. Nevertheless, $\mathbf{N}_m$ is not constrained by the labels, allowing flexible codeword assignments for multi-label representation.

Label-Guided Geometric Structure Preservation. Although semantic alignment and implicit codebook learning provide discriminative compact representations, the semantic neighborhood relationships among samples may still be distorted during quantization. To address this issue, we preserve pairwise semantic relationships using the normalized label matrix $\bar{\mathbf{L}}$. Conceptually, the label-guided semantic affinity can be written as $\mathbf{S} = \bar{\mathbf{L}}\bar{\mathbf{L}}^\top \in \mathbb{R}^{N\times N}$. However, $\mathbf{S}$ is not explicitly materialized in our implementation. Instead, the matrix products involving $\mathbf{S}$ are computed through the low-dimensional label representation $\bar{\mathbf{L}}$ by exploiting the associativity of matrix multiplication. This avoids the storage cost of an explicit $N \times N$ graph while preserving global semantic relationships.

Accordingly, we enforce the semantic affinities induced by the learned representations to be consistent with the label-guided semantic structure:

$$\sum_{p=1}^{P} \|\bar{\mathbf{L}}\bar{\mathbf{L}}^\top - \mathbf{X}_p^\top\mathbf{Z}_p^\top\mathbf{Q}_p\mathbf{S}_p\bar{\mathbf{V}}^\top\|_{\mathrm{F}}^2. \tag{5}$$

This term explicitly aligns the label-induced semantic affinity with the cross-space affinity between continuous modality-specific representations and quantized representations. Therefore, the learned discrete codes are encouraged to preserve the same semantic neighborhood structure as the supervisory labels.

Overall Objective. By combining Eq.(1), Eq.(3), and Eq.(5), the overall objective function can be defined as follows:

$$\begin{aligned} \min_{\Theta} \mathcal{J} = &\sum_{p=1}^{P} \|\mathbf{Z}_p\mathbf{X}_p - \mathbf{Q}_p\mathbf{H}\mathbf{S}_p\mathbf{V}^\top\|_{\mathrm{F}}^2 + \alpha \sum_{m=1}^{M} \|\mathbf{V}_m\mathbf{W}_m + \mathbf{1}_N\mathbf{t}^\top - \mathbf{Y}\|_{\mathrm{F}}^2 \\ &+ \beta \sum_{p=1}^{P} \|\bar{\mathbf{L}}\bar{\mathbf{L}}^\top - \mathbf{X}_p^\top\mathbf{Z}_p^\top\mathbf{Q}_p\mathbf{S}_p\bar{\mathbf{V}}^\top\|_{\mathrm{F}}^2 + \mu\|\mathbf{H}\|_{\mathrm{F}}^2 \\ s.t. \quad &\mathbf{Q}_p^\top\mathbf{Q}_p = \mathbf{I}_r,\ \mathbf{1}_c^T\mathbf{N}_m = \mathbf{1}_N^\top,\ \mathbf{D} \geq 0 \quad \forall m \in \{1,\ldots,M\} \quad \forall p \in \{1,\ldots,P\} \end{aligned} \tag{6}$$

where $\Theta = \{\mathbf{Z}_p, \mathbf{Q}_p, \mathbf{H}, \mathbf{S}_p, \mathbf{V}_m, \mathbf{W}_m, \mathbf{N}_m, \mathbf{t}, \mathbf{D}\}$, and α, β, and μ are non-negative trade-off hyperparameters governing the importance of label regression, geometric preservation, and shared semantic basis regularization, respectively.

2.3 Optimization

Since jointly optimizing all variables in Eq. (6) is challenging, we develop an alternating optimization algorithm. By fixing all but one variable at each iteration, the objective function can be decomposed into nine independent subproblems, which are solved in a cyclic manner until convergence. For a detailed derivation, see the Appendix B.

Step 1: Update $\mathbf{Z}_p$ with $\partial\mathcal{J}/\partial\mathbf{Z}_p = 0$, we have

$$\mathbf{Z}_p = \left(\mathbf{I}_r + \beta\mathbf{B}_2\mathbf{B}_2^\top\right)^{-1}\left(\mathbf{B}_1 + \beta(\mathbf{B}_2\bar{\mathbf{L}})\bar{\mathbf{L}}^\top\right)\mathbf{X}_p^\top\left(\mathbf{X}_p\mathbf{X}_p^\top\right)^{-1}, \tag{7}$$

where $\mathbf{B}_1 = \mathbf{Q}_p\mathbf{H}\mathbf{S}_p\mathbf{V}^\mathsf{T} \in \mathbb{R}^{r\times N}$, $\mathbf{B}_2 = \mathbf{Q}_p\mathbf{S}_p\bar{\mathbf{V}}^\mathsf{T} \in \mathbb{R}^{r\times N}$, and $\mathbf{I}_r \in \mathbb{R}^{r\times r}$ is an identity matrix.

Step 2: Update $\mathbf{Q}_p$ with others fixed, the subproblem is

$$\begin{aligned} \min_{\mathbf{Q}_p} \|\hat{\mathbf{X}}_p - \mathbf{Q}_p\mathbf{H}\mathbf{S}_p\mathbf{V}^\mathsf{T}\|_\mathrm{F}^2 + \beta\|\bar{\mathbf{L}}\bar{\mathbf{L}}^\mathsf{T} - \hat{\mathbf{X}}_p^\mathsf{T}\mathbf{Q}_p\mathbf{S}_p\bar{\mathbf{V}}^\mathsf{T}\|_\mathrm{F}^2, \\ s.t. \quad \mathbf{Q}_p^\mathsf{T}\mathbf{Q}_p = \mathbf{I}_r, \end{aligned} \tag{8}$$

where $\hat{\mathbf{X}}_p = \mathbf{Z}_p\mathbf{X}_p$. We optimize $\mathbf{Q}_p$ on the Stiefel manifold to satisfy the orthogonality constraint Wen & Yin (2013). Let $\mathbf{Q}_p(i)$ denote the solution at the i-th iteration. The update at iteration $i+1$ is given by

$$\mathbf{Q}_p(i+1) = \left(\mathbf{I} + \frac{\kappa}{2}\boldsymbol{\nabla}\right)^{-1}\left(\mathbf{I} - \frac{\kappa}{2}\boldsymbol{\nabla}\right)\mathbf{Q}_p(i), \tag{9}$$

where κ is a proper step size determined by nonmonotonic line search. $\boldsymbol{\nabla} = \mathbf{G}(\mathbf{Q}_p)\mathbf{Q}_p(i)^\mathsf{T} - \mathbf{Q}_p(i)\mathbf{G}(\mathbf{Q}_p)^\mathsf{T}$ is a defined skew-symmetric matrix with $\mathbf{G}(\mathbf{Q}_p)$ calculated as

$$\mathbf{G}(\mathbf{Q}_p) = 2(\mathbf{B}_1 - \hat{\mathbf{X}}_p)\mathbf{V}\mathbf{S}_p\mathbf{H}^\mathsf{T} + 2\beta\hat{\mathbf{X}}_p(\hat{\mathbf{X}}_p^\mathsf{T}\mathbf{B}_2 - \bar{\mathbf{L}}\bar{\mathbf{L}}^\mathsf{T})\bar{\mathbf{V}}\mathbf{S}_p. \tag{10}$$

Step 3: Update $\mathbf{H}$ with $\partial\mathcal{J}/\partial\mathbf{H} = 0$, we have

$$\mathbf{H} = \left(\sum_{p=1}^{P}\mathbf{Q}_p^\mathsf{T}\hat{\mathbf{X}}_p\mathbf{V}\mathbf{S}_p\right)\left(\sum_{p=1}^{P}\mathbf{S}_p\mathbf{V}^\mathsf{T}\mathbf{V}\mathbf{S}_p + \mu\mathbf{I}_r\right)^{-1}. \tag{11}$$

Step 4: To optimize the block variable $\mathbf{V}_m$, we define $\mathbf{B}_p = \mathbf{S}_p\mathbf{H}^\mathsf{T}\mathbf{Q}_p^\mathsf{T}$, and divide it according to the block structure of $\mathbf{V}$ as $\mathbf{B}_p = [\mathbf{B}_{p,1}; \mathbf{B}_{p,2}; \ldots; \mathbf{B}_{p,M}]$. Then, the matrix product can be decomposed as $\mathbf{V}\mathbf{B}_p = \sum_{j=1}^{M}\mathbf{V}_j\mathbf{B}_{p,j} = \mathbf{V}_m\mathbf{B}_{p,m} + \sum_{j\neq m}\mathbf{V}_j\mathbf{B}_{p,j}$. Substituting the above expression into the objective function and discarding the terms independent of $\mathbf{V}_m$, the subproblem is formulated as:

$$\min_{\mathbf{V}_m}\mathcal{J}(\mathbf{V}_m) = \sum_{p=1}^{P}\|(\hat{\mathbf{X}}_p^\mathsf{T} - \sum_{j\neq m}\mathbf{V}_j\mathbf{B}_{p,j}) - \mathbf{V}_m\mathbf{B}_{p,m}\|_\mathrm{F}^2 + \alpha\|\mathbf{V}_m\mathbf{W}_m + \mathbf{1}_N\mathbf{t}^\mathsf{T} - \mathbf{Y}\|_\mathrm{F}^2. \tag{12}$$

Update $\mathbf{V}_m$ with $\partial\mathcal{J}(\mathbf{V}_m)/\partial\mathbf{V}_m = 0$, we obtain

$$\mathbf{V}_m = \left(\sum_{p=1}^{P}\left(\hat{\mathbf{X}}_p^\mathsf{T} - \sum_{j\neq m}\mathbf{V}_j\mathbf{B}_{p,j}\right)\mathbf{B}_{p,m}^\mathsf{T} + \alpha\bar{\mathbf{Y}}\mathbf{W}_m^\mathsf{T}\right)\left(\sum_{p=1}^{P}\mathbf{B}_{p,m}\mathbf{B}_{p,m}^\mathsf{T} + \alpha\mathbf{W}_m\mathbf{W}_m^\mathsf{T}\right)^{-1}. \tag{13}$$

where $\bar{\mathbf{Y}} = \mathbf{Y} - \mathbf{1}_N\mathbf{t}^\mathsf{T}$. Then, $\mathbf{V}$ is constructed by column-wise concatenation, i.e., $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_M]$.

Step 5: Update $\mathbf{t}$ with $\partial\mathcal{J}/\partial\mathbf{t} = 0$, we have

$$\mathbf{t} = \frac{1}{MN}\sum_{m=1}^{M}(\mathbf{Y} - \mathbf{V}_m\mathbf{W}_m)^\mathsf{T}\mathbf{1}_N. \tag{14}$$

Step 6: For $\mathbf{D}$, setting $\partial\mathcal{J}/\partial\mathbf{D} = 0$ gives

$$-2\alpha\sum_{m=1}^{M}\mathbf{C}\odot(\mathbf{V}_m\mathbf{W}_m + \mathbf{1}_N\mathbf{t}^\mathsf{T} - \mathbf{L} - \mathbf{C}\odot\mathbf{D}) = 0. \tag{15}$$

Exploiting $\mathbf{C}\odot\mathbf{C} = \mathbf{1}$ and enforcing $\mathbf{D} \geq 0$, the closed-form solution is

$$\mathbf{D} = \max\left(\frac{1}{M}\sum_{m=1}^{M}\mathbf{C}\odot\left(\mathbf{V}_m\mathbf{W}_m + \mathbf{1}_N\mathbf{t}^\mathsf{T} - \mathbf{L}\right),\ 0\right). \tag{16}$$

Step 7: Updating $\mathbf{S}_p$. Since $\mathbf{S}_p = \mathrm{diag}(\mathbf{s}_p)$, the subproblem can be written as

$$\min_{\mathbf{s}_p}\mathcal{J}(\mathbf{s}_p) = \mathbf{s}_p^\mathsf{T}\boldsymbol{\Gamma}_p\mathbf{s}_p - 2\mathbf{s}_p^\mathsf{T}\boldsymbol{\gamma}_p, \tag{17}$$

where $\mathbf{\Gamma}_p = (\mathbf{H}^\top\mathbf{H}) \odot (\mathbf{V}^\top\mathbf{V}) + \beta(\mathbf{Q}_p^\top\hat{\mathbf{X}}_p\hat{\mathbf{X}}_p^\top\mathbf{Q}_p) \odot (\bar{\mathbf{V}}^\top\bar{\mathbf{V}})$ and $\boldsymbol{\gamma}_p = \operatorname{diag}(\mathbf{H}^\top\mathbf{Q}_p^\top\hat{\mathbf{X}}_p\mathbf{V}) + \beta \operatorname{diag}((\mathbf{Q}_p^\top\hat{\mathbf{X}}_p\bar{\mathbf{L}})(\bar{\mathbf{L}}^\top\bar{\mathbf{V}}))$. The closed-form solution is

$$\mathbf{s}_p = \mathbf{\Gamma}_p^{-1}\boldsymbol{\gamma}_p \quad \Longrightarrow \quad \mathbf{S}_p = \operatorname{diag}(\mathbf{s}_p). \tag{18}$$

Step 8: Updating $\mathbf{W}_m$. Since $\mathbf{W}_m$ contains only c category-aware codewords, the codebook update can be efficiently performed column by column. The subproblem is

$$\min_{\mathbf{W}_m} \mathcal{J}(\mathbf{W}_m) = \alpha\|\mathbf{V}_m\mathbf{W}_m - \bar{\mathbf{Y}}\|_F^2 + \beta\sum_{p=1}^{P}\|\mathbf{E}_{p,-m} - \mathbf{A}_{p,m}\mathbf{W}_m\mathbf{N}_m\|_F^2, \tag{19}$$

where $\mathbf{E}_{p,-m} = \bar{\mathbf{L}}\bar{\mathbf{L}}^\top - \sum_{j\neq m}\mathbf{A}_{p,j}\mathbf{W}_j\mathbf{N}_j$ and $\mathbf{A}_p = \hat{\mathbf{X}}_p^\top\mathbf{Q}_p\mathbf{S}_p = [\mathbf{A}_{p,1}, \ldots, \mathbf{A}_{p,M}]$.

Setting $\partial\mathcal{J}/\partial\mathbf{W}_m = 0$ gives

$$\mathbf{\Phi}_m\mathbf{W}_m + \mathbf{\Psi}_m\mathbf{W}_m\mathbf{\Lambda}_m = \mathbf{\Omega}_m, \tag{20}$$

where $\mathbf{\Phi}_m = \alpha\mathbf{V}_m^\top\mathbf{V}_m$, $\mathbf{\Psi}_m = \beta\sum_{p=1}^{P}\mathbf{A}_{p,m}^\top\mathbf{A}_{p,m}$, $\mathbf{\Lambda}_m = \mathbf{N}_m\mathbf{N}_m^\top$, and $\mathbf{\Omega}_m = \alpha\mathbf{V}_m^\top\bar{\mathbf{Y}} + \beta\sum_{p=1}^{P}\mathbf{A}_{p,m}^\top\mathbf{E}_{p,-m}\mathbf{N}_m^\top$. Since $\mathbf{\Lambda}_m = \operatorname{diag}(c_1, \ldots, c_c)$, the columns of $\mathbf{W}_m$ can be updated independently as

$$\mathbf{W}_{m,i} = (\mathbf{\Phi}_m + c_i\mathbf{\Psi}_m + \epsilon\mathbf{I}_d)^{-1}\mathbf{\Omega}_{m,i}, \quad i = 1, \ldots, c, \tag{21}$$

where ϵ is a very small positive constant.

Step 9: Updating $\mathbf{N}_m$. The subproblem is

$$\begin{aligned} \min_{\mathbf{N}_m} \ & \sum_{p=1}^{P}\|\mathbf{E}_{p,-m} - \mathbf{A}_{p,m}\mathbf{W}_m\mathbf{N}_m\|_F^2 \\ \text{s.t. } & \mathbf{1}_c^\top\mathbf{N}_m = \mathbf{1}_N^\top. \end{aligned} \tag{22}$$

After expanding the Frobenius norm and removing constant terms, the problem reduces to

$$\min_{\mathbf{N}_m} \operatorname{Tr}\left(\mathbf{\Delta}_m^\top\mathbf{N}_m\right), \tag{23}$$

where $\mathbf{\Delta}_m = \operatorname{diag}(\mathbf{G}_m)\mathbf{1}_N^\top - 2\mathbf{R}_m$, where $\mathbf{G}_m = \sum_{p=1}^{P}\mathbf{W}_m^\top\mathbf{A}_{p,m}^\top\mathbf{A}_{p,m}\mathbf{W}_m$ and $\mathbf{R}_m = \sum_{p=1}^{P}\mathbf{W}_m^\top\mathbf{A}_{p,m}^\top\mathbf{E}_{p,-m}$. Thus, each sample can be assigned independently:

$$j_i^* = \arg\min_{j\in\{1,\ldots,c\}}(\mathbf{\Delta}_m)_{j,i}, \quad i = 1, \ldots, N. \tag{24}$$

The assignment matrix is then updated by setting $(\mathbf{N}_m)_{j_i^*,i} = 1$ and all remaining entries to zero.

2.4 Out-of-sample Extension and Retrieval

For an unseen query $\mathbf{x}_q^{(p)}$ from the p-th modality, we first project it into the learned latent semantic space:

$$\mathbf{b}_p^q = \mathbf{x}_p^{q\top}\mathbf{Z}_p^\top\mathbf{Q}_p\mathbf{S}_p. \tag{25}$$

The database samples are represented by the learned $\bar{\mathbf{V}}$. Following the asymmetric distance computation strategy in PQ Jegou et al. (2010), we compute the Euclidean distance between the continuous query representation and the codewords selected by database samples. This enables efficient CMR without re-optimizing the model for new queries.

2.5 Complexity Analysis

To demonstrate the high efficiency of P2PQ, we conduct a detailed computational complexity analysis. The main complexity of each iteration mainly includes $\mathcal{O}(N(d_p^2 + rd_p + r^2 + rc) + d_p^3 + r^3)$ for updating Eq. (7), $\mathcal{O}(Pr^2N + Pr^3)$ for updating Eq. (11), $\mathcal{O}(PNr^2 + PNrd + Ncd + Pd^2r + r^3)$ for updating Eq. (13), $\mathcal{O}(MNdc)$ for updating Eq. (14) and Eq. (16), $\mathcal{O}(Nr^2 + rNc + r^3)$ for updating Eq. (18), $\mathcal{O}(PNrd + PNdc + PNc^2 + Prdc + Prc^2 + PMNc + cd^3)$ for updating Eq. (21), and $\mathcal{O}(PNrd + Prdc + PMNdc + Pcd^2 + Pc^2d + PNdc + Nc)$ for updating Eq. (24). Furthermore, according to Wen & Yin (2013); Huang et al. (2024), the computational complexity of solving $\mathbf{Q}_p$ is approximately $\mathcal{O}(\kappa r^3)$. Since $d_p, r, c, d, P, M \ll N$, we can observe the overall complexity approximates to $\mathcal{O}(N)$, i.e., linear to the size of the training set.

3 EXPERIMENTS

3.1 DATASETS AND EVALUATION PROTOCOL

We evaluate P2PQ on three widely used CMR datasets, i.e., MIRFLICKR-25K Huiskes & Lew (2008), IAPRTC-12 Escalante et al. (2010), and NUS-WIDE Chua et al. (2009). A detailed description of the datasets can be found in Appendix C. We evaluate two standard CMR tasks: image-to-text retrieval (I2T) and text-to-image retrieval (T2I), and report mean Average Precision (mAP) and precision-recall (PR) curves.

3.2 BASELINES AND IMPLEMENTATION DETAILS

Since P2PQ focuses on shallow compact representation learning, we compare it with 12 representative shallow cross-modal hashing methods, including BATCH Wang et al. (2021), DAH Zhang et al. (2023b), WASH Zhang et al. (2023a), ALECH Li et al. (2023), DCGH Sun et al. (2024b), EDH Huang et al. (2024), DSFH Liu et al. (2024), DEH-IL Zhang et al. (2025), HPsEH Wong et al. (2025), CSMH Li et al. (2025b), MPSH Zheng et al. (2026), and RCMH Jiang et al. (2026). All baselines are implemented using the optimal parameter settings reported in their original papers. For P2PQ, the latent dimension is defined as $r = Md$, where M is the number of subspaces and d is the dimensionality of each subspace. Since the number of categories is no greater than 256 on all datasets, each subspace index is stored using 8 bits. Accordingly, for total code lengths of 16, 32, 64, and 128 bits, the number of subspaces M is set to 2, 4, 8, and 16, respectively. For NUS-WIDE, d is set to 10 under 64- and 128-bit settings and 30 under 16- and 32-bit settings. For MIRFLICKR-25K and IAPRTC-12, d is fixed to 20 and 30, respectively. The trade-off parameters $\{\alpha, \beta, \mu\}$ are set to $\{10^5, 1, 10^{-2}\}$, $\{10^5, 1, 10^{-4}\}$, and $\{10^3, 10, 10^2\}$ for NUS-WIDE, MIRFLICKR-25K, and IAPRTC-12, respectively. The numbers of image and text anchors are set to $\{1150, 1250\}$, $\{800, 1400\}$, and $\{1000, 1000\}$ for three datasets. In addition, the iteration number is set to 15. All experiments are conducted on a laptop equipped with an Intel Core i7-14650HX CPU and 24 GB RAM.

Table 1: Mean Average Precision (mAP) comparison of P2PQ and baseline methods on MIRFLICKR-25K, IAPRTC-12, and NUS-WIDE under different code lengths. The best results are shown in bold and the second-best results are underlined.

Task	Method	MIRFLICKR-25K				IAPRTC-12				NUS-WIDE			
		16	32	64	128	16	32	64	128	16	32	64	128
I2T	BATCH (TKDE'21)	0.7375	0.7438	0.7452	0.7494	0.4799	0.5031	0.5246	0.5372	0.6274	0.6505	0.6689	0.6685
	DAH (TKDE'23)	0.7191	0.7281	0.7329	0.7361	0.4736	0.5161	0.5403	0.5636	0.6615	0.6776	0.6880	0.6923
	WASH (TKDE'23)	0.7344	0.7422	0.7466	0.7492	0.4844	0.5100	0.5345	0.5461	0.6581	0.6660	0.6744	0.6777
	ALECH (TKDE'23)	0.7308	0.7384	0.7421	0.7438	0.4798	0.5076	0.5212	0.5341	0.6598	0.6724	0.6745	0.6824
	DCGH (MM'24)	0.7539	0.7586	0.7617	0.7635	0.5234	0.5476	0.5635	0.5720	0.6702	0.6794	0.6822	0.6844
	EDH (TSMCS'24)	0.7173	0.7438	0.7549	0.7627	0.4706	0.4972	0.5113	0.5377	0.6507	0.6550	0.6625	0.6707
	DSFH (IJCAI'24)	0.7482	0.7555	0.7616	0.7638	0.5267	0.5466	0.5639	0.5745	0.6689	0.6791	0.6825	0.6866
	DEH-IL (TOMM'25)	0.7294	0.7353	0.7409	0.7435	0.4888	0.5155	0.5335	0.5455	0.6602	0.6735	0.6824	0.6857
	HPsEH (TMM'25)	0.7265	0.7359	0.7404	0.7433	0.4875	0.5071	0.5242	0.5402	0.6602	0.6716	0.6781	0.6823
	CSMH (TKDE'25)	0.7393	0.7526	0.7559	0.7592	0.4805	0.5040	0.5247	0.5378	0.6789	0.6832	0.6866	0.6894
	MPSH (TCSVT'26)	0.7530	0.7587	0.7618	0.7632	0.5089	0.5345	0.5503	0.5613	0.6547	0.6582	0.6648	0.6676
	RCMH (TPAMI'26)	0.7373	0.7555	0.7664	0.7728	0.5113	0.5348	0.5506	0.5640	0.6661	0.6778	0.6861	0.6915
	P2PQ (Ours)	**0.7932**	**0.7908**	**0.7891**	**0.7839**	**0.5916**	**0.5951**	**0.5989**	**0.6001**	**0.7076**	**0.7173**	**0.7148**	**0.7173**
T2I	BATCH (TKDE'21)	0.8217	0.8274	0.8319	0.8376	0.5758	0.6185	0.6504	0.6683	0.7596	0.7782	0.7815	0.7837
	DAH (TKDE'23)	0.8151	0.8267	0.8392	0.8384	0.5505	0.5941	0.6362	0.6622	0.7734	0.7977	0.8072	0.8142
	WASH (TKDE'23)	0.8141	0.8279	0.8325	0.8345	0.5803	0.6176	0.6508	0.6660	0.7827	0.7915	0.8039	0.8082
	ALECH (TKDE'23)	0.8282	0.8392	0.8438	0.8477	0.5702	0.6169	0.6456	0.6615	0.7841	0.7928	0.8014	0.8112
	DCGH (MM'24)	0.8297	0.8389	0.8428	0.8451	0.6064	0.6453	0.6737	0.6893	0.8028	0.8127	0.8179	0.8187
	EDH (TSMCS'24)	0.7983	0.8306	0.8440	0.8527	0.5463	0.6051	0.6366	0.6605	0.7803	0.8020	0.8098	0.8276
	DSFH (IJCAI'24)	0.8283	0.8385	0.8455	0.8501	0.6086	0.6467	0.6730	0.6892	0.8041	0.8229	0.8301	0.8346
	DEH-IL (TOMM'25)	0.8284	0.8362	0.8428	0.8448	0.5781	0.6230	0.6524	0.6698	0.7937	0.8058	0.8139	0.8171
	HPsEH (TMM'25)	0.8194	0.8299	0.8350	0.8376	0.5808	0.6193	0.6544	0.6779	0.7885	0.8047	0.8095	0.8137
	CSMH (TKDE'25)	0.8229	0.8285	0.8328	0.8407	0.5762	0.6208	0.6583	0.6765	0.7997	0.8054	0.8099	0.8163
	MPSH (TCSVT'26)	0.8406	0.8504	0.8552	0.8575	0.5999	0.6399	0.6684	0.6847	0.7880	0.7992	0.8061	0.8090
	RCMH (TPAMI'26)	0.8139	0.8210	0.8286	0.8288	0.6064	0.6514	0.6852	0.7083	0.8025	0.8176	0.8285	0.8352
	P2PQ (Ours)	**0.8814**	**0.8858**	**0.8862**	**0.8836**	**0.7020**	**0.7124**	**0.7287**	**0.7375**	**0.8308**	**0.8548**	**0.8556**	**0.8594**

3.3 OVERALL COMPARISON

Table 1 reports the mAP results of all compared methods on the three datasets under different code lengths. P2PQ achieves the best performance in most settings for both I2T and T2I retrieval, demon-

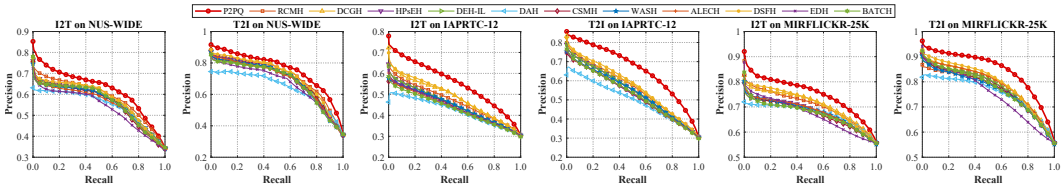


Figure 2: PR curves of P2PQ and the baseline methods on three datasets under 16-bit settings.

strating the effectiveness of the proposed unified framework. The improvements are particularly clear under short code lengths, such as 16 and 32 bits, indicating that P2PQ can learn compact yet semantically discriminative quantized representations. This advantage stems from the joint effect of semantic alignment, category-aware implicit codebook learning, and label-guided geometric preservation. Specifically, semantic alignment reduces heterogeneous modality discrepancies, the implicit codebooks couple quantization with semantic regression, and geometric preservation maintains semantic relationship consistency between continuous and quantized representations. On the large-scale NUS-WIDE dataset, P2PQ also maintains superior performance, showing its scalability for large-scale CMR. Fig. 2 presents the PR curves under the 16bits setting, where P2PQ consistently achieves higher precision across a wide range of recall values. Due to the page limit, we present additional results in Appendix D, including the PR curves under the 64bits setting and the P@N curves under both the 16- and 64-bit settings, which further demonstrate the consistent superiority of P2PQ.

3.4 Ablation Study

To evaluate the contribution of each component, we construct three variants: P2PQ-1 removes the RBF kernel mapping, P2PQ-2 only retains the label-guided geometric preservation term in Eq. (5), removing both semantic alignment and semantic regression, and P2PQ-3 removes Eq. (5) and encodes database samples by the minimum Euclidean distance in each subspace. As shown in Table 2, all variants perform worse than the full model. The degradation of P2PQ-1 verifies the benefit of nonlinear feature mapping. The poor performance of P2PQ-2 indicates that label-guided geometric preservation alone is insufficient, demonstrating the importance of jointly incorporating semantic alignment and regression into the unified framework. The decline of P2PQ-3 further confirms the importance of label-guided geometric preservation. These results demonstrate that the three major components work jointly to produce effective quantized representations.

Table 2: The Performance of P2PQ and its Variants on the Three Datasets with Various Code Lengths

Task	Method	MIRFLICKR-25K				IAPRTC-12				NUS-WIDE			
		16	32	64	128	16	32	64	128	16	32	64	128
I2T	P2PQ-1	0.7761	0.7738	0.7733	0.7675	0.5646	0.5692	0.5688	0.5685	0.6937	0.7026	0.6991	0.7010
	P2PQ-2	0.6112	0.6055	0.6114	0.6125	0.3433	0.3050	0.3035	0.3325	0.4369	0.4339	0.4505	0.3832
	P2PQ-3	0.5708	0.5936	0.5947	0.5918	0.2920	0.2859	0.3110	0.3175	0.4045	0.4218	0.4333	0.4230
	P2PQ	**0.7932**	**0.7908**	**0.7891**	**0.7839**	**0.5916**	**0.5951**	**0.5989**	**0.6001**	**0.7076**	**0.7173**	**0.7148**	**0.7173**
T2I	P2PQ-1	0.8347	0.8381	0.8384	0.8311	0.6726	0.6960	0.6992	0.7072	0.8126	0.8406	0.8421	0.8459
	P2PQ-2	0.6127	0.6072	0.6124	0.6132	0.3433	0.3050	0.3035	0.3325	0.4406	0.4383	0.4529	0.3864
	P2PQ-3	0.5749	0.5826	0.5758	0.5718	0.2938	0.3013	0.3151	0.3128	0.3749	0.3986	0.4094	0.4068
	P2PQ	**0.8814**	**0.8858**	**0.8862**	**0.8836**	**0.7020**	**0.7124**	**0.7287**	**0.7375**	**0.8308**	**0.8548**	**0.8556**	**0.8594**

3.5 Parameter Sensitivity Analysis

To investigate the robustness of P2PQ, we evaluate the influence of α, β, d, and μ under the 16 bits setting. As shown in Fig. 3, P2PQ maintains relatively stable performance over a broad range of α and β, while overly small or large values lead to suboptimal results. Since α controls semantic regression and β controls label-guided geometric preservation, the best performance is achieved when these two terms are properly balanced. In particular, a too small α weakens the discriminative role of the implicit codebooks, whereas a too small β insufficiently preserves the semantic relationship consistency between continuous and quantized representations. Fig. 4(a) shows that the effect of the subspace dimension d is dataset-dependent. P2PQ is relatively stable on MIRFLICKR-25K, while larger fluctuations are observed on IAPRTC-12 and NUS-WIDE, suggesting that an appropriate subspace partition is important for balancing representation capacity and quantization quality. Fig. 4(b)

further shows that P2PQ is generally insensitive to μ over several orders of magnitude. Overall, these results demonstrate that P2PQ is robust to parameter variations, while the balance between semantic regression and geometric preservation remains important.

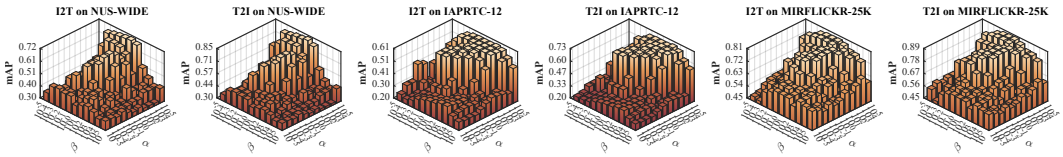
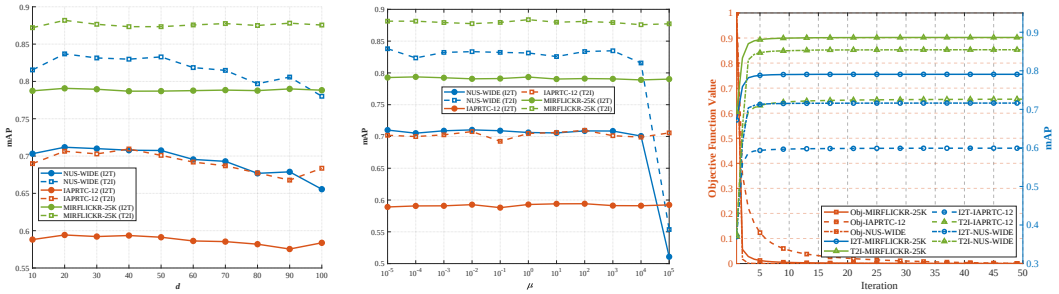


Figure 3: mAP versus parameter α and β.



(a) d (b) μ (c) Convergence curves

Figure 4: mAP versus parameter d and μ under 16-bit setting, and convergence curves of P2PQ on three datasets using 32-bit code length.

3.6 Convergence Analysis

Fig. 4(c) shows the convergence behavior of P2PQ on the three datasets, where the objective values are normalized to [0,1] for better visualization. The objective value decreases consistently during optimization, indicating the stability of the proposed alternating optimization algorithm. Meanwhile, the I2T and T2I mAP values become stable after several iterations. Although the objective value on IAPRTC-12 decreases more gradually, its retrieval performance converges earlier, showing that the learned representations can reach stable retrieval accuracy before the objective value fully stabilizes.

Due to space limitations, additional experimental results, including the comparison with deep-feature-based methods and the detailed training time analysis, are provided in Appendix D.

4 Conclusion

In this paper, we proposed P2PQ, a unified product quantization framework for cross-modal retrieval. By learning category-aware implicit codebooks whose size is determined by the number of semantic categories, P2PQ allows the codebooks to simultaneously act as quantization prototypes and semantic regression weights, thus coupling discrete quantization with semantic discrimination while reducing the number of codewords. Moreover, multimodal semantic alignment and label-guided geometric preservation are jointly integrated into the same objective to improve cross-modal consistency and preserve semantic relationships between continuous and quantized representations. Extensive experiments on three benchmark datasets demonstrate that P2PQ consistently outperforms representative shallow cross-modal retrieval methods under various code lengths, validating its effectiveness and scalability for large-scale retrieval.

AI use statement

In this work, we used generative AI tools solely for language editing and polishing to improve the clarity, grammar, and readability of the manuscript. Generative AI tools were not used to generate

research ideas, formulate the proposed method, conduct experiments, analyze experimental results, or derive scientific conclusions.

All AI-assisted edits were carefully reviewed and revised by the authors to ensure technical accuracy, consistency with the original intended meaning, and appropriate academic presentation. We take full responsibility for the final content of this work, including all text and claims presented in the manuscript.

Reproducibility statement

We have taken several steps to facilitate the reproducibility of our work. The main paper describes the proposed method, experimental settings, datasets, evaluation metrics, and comparisons with competing approaches. Additional implementation details, hyperparameter settings, and supplementary experimental results are provided in the appendix and supplementary materials. We also report the configurations used in the parameter sensitivity analyses and provide sufficient details of the experimental protocol to support independent reproduction of the reported results.

References

Yue Cao, Mingsheng Long, Jianmin Wang, and Shichen Liu. Collective deep quantization for efficient cross-modal retrieval. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3974–3980, 2017.

Bin Chen, Yan Feng, Tao Dai, Jiawang Bai, Yong Jiang, Shu-Tao Xia, and Xuan Wang. Adversarial examples generation for deep product quantization networks on image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(2):1388–1404, 2022.

Haipeng Chen, Yu Liu, Xun Yang, Yuheng Liang, and Yingda Lyu. Causality-aligned semantic recovery for incomplete cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 20050–20058, 2026.

Zhen-Duo Chen, Chuan-Xiang Li, Xin Luo, Liqiang Nie, Wei Zhang, and Xin-Shun Xu. Scratch: A scalable discrete matrix factorization hashing framework for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 30(7):2262–2275, 2020.

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 ACM International Conference on Image and Video Retrieval (CIVR)*, 2009.

Xinfeng Dong, Li Liu, Lei Zhu, Liqiang Nie, and Huaxiang Zhang. Adversarial graph convolutional network for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 32(3):1634–1645, 2022.

Hugo Jair Escalante, Carlos A. Hernández, Jesus A. Gonzalez, A. López-López, Manuel Montes, Eduardo F. Morales, L. Enrique Sucar, Luis Villaseñor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428, 2010.

Lianli Gao, Xiaosu Zhu, Jingkuan Song, Zhou Zhao, and Heng Tao Shen. Beyond product quantization: deep progressive quantization for image retrieval. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 723–729, 2019.

Shuo Han, Qibing Qin, Wenfeng Zhang, and Lei Huang. Deep uncertainty-aware probabilistic hashing for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)*, 22(2):1–23, 2026.

Zhikai Hu, Yiu-Ming Cheung, Mengke Li, and Weichao Lan. Cross-modal hashing method with properties of hamming space: A new perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(12):7636–7650, 2024.

Junfan Huang, Peipei Kang, Xiaozhao Fang, Na Han, Shengli Xie, and Hongbo Gao. Efficient discriminative hashing for cross-modal retrieval. *IEEE Transactions on Systems, Man, and Cybernetics: Systems (TSMCS)*, 54(6):3865–3878, 2024.

Iris A. M. Huijben, Matthijs Douze, Matthew Muckley, Ruud J. G. Van Sloun, and Jakob Verbeek. Residual quantization with implicit neural codebooks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 20682–20699, 2024.

Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval (MIR)*, pp. 39–43, 2008.

Yadong Huo, Qibing Qin, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep hierarchy-aware proxy hashing with self-paced learning for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 36(11):5926–5939, 2024.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(1):117–128, 2010.

Kaihang Jiang, Wai Keung Wong, Xiaozhao Fang, Weijun Sun, Guoxu Zhou, Shengli Xie, and Xiaochun Cao. An efficient regenerated cross-modal hashing: Improving existing hash codes with the arbitrary length. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–18, 2026. doi: 10.1109/TPAMI.2026.3688816.

Huaxiong Li, Chao Zhang, Xiuyi Jia, Yang Gao, and Chunlin Chen. Adaptive label correlation based asymmetric discrete hashing for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 35(2):1185–1199, 2023.

Jiaxing Li, Lin Jiang, Zeqi Ma, Kaihang Jiang, Xiaozhao Fang, and Jie Wen. Lightweight contrastive distilled hashing for online cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4779–4787, 2025a.

Jiaxing Li, Wai Keung Wong, Lin Jiang, Kaihang Jiang, Xiaozhao Fang, Shengli Xie, and Jie Wen. Collaboratively semantic alignment and metric learning for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 37(5):2311–2328, 2025b.

Kaiming Liu, Yunhong Gong, Yu Cao, Zhenwen Ren, Dezhong Peng, and Yuan Sun. Dual semantic fusion hashing for multi-label cross-modal retrieval. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4569–4577, 2024.

Yuhao Liu, Yanbo Zhang, Hao Fu, and Guanghua Gu. Adaptive asymmetric online hashing for cross-modal retrieval. In *Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR)*, pp. 908–916, 2025.

Mingsheng Long, Yue Cao, Jianmin Wang, and Philip S. Yu. Composite correlation quantization for efficient multimodal retrieval. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 579–588, 2016.

Ruitao Pu, Yuan Sun, Yang Qin, Zhenwen Ren, Xiaomin Song, Huiming Zheng, and Dezhong Peng. Robust self-paced hashing for cross-modal retrieval with noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 19969–19977, 2025.

Jingkuan Song, Xiaosu Zhu, Lianli Gao, Xin-Shun Xu, Wu Liu, and Heng Tao Shen. Deep recurrent quantization for generating sequential binary codes. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 912–918, 2019.

Chao Su, Likang Peng, Yuan Sun, Dezhong Peng, Xi Peng, and Xu Wang. Neighbor-aware contrastive disambiguation for cross-modal hashing with redundant annotations. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:108581–108603, 2026.

Hao Sun, Yadong Huo, Qibing Qin, Wenfeng Zhang, and Lei Huang. Intra-class distribution-guided generative hashing with neighbor refinement for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2671–2681, 2026.

Yuan Sun, Jian Dai, Zhenwen Ren, Yingke Chen, Dezhong Peng, and Peng Hu. Dual self-paced cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 15184–15192, 2024a.

Yuan Sun, Kaiming Liu, Yongxiang Li, Zhenwen Ren, Jian Dai, and Dezhong Peng. Distribution consistency guided hashing for cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, pp. 5623–5632, 2024b.

Jun Tang, Ke Wang, and Ling Shao. Supervised matrix factorization hashing for cross-modal retrieval. *IEEE Transactions on Image Processing (TIP)*, 25(7):3157–3166, 2016.

Lu Wang, Masoumeh Zareapoor, Jie Yang, and Zhonglong Zheng. Asymmetric correlation quantization hashing for cross-modal retrieval. *IEEE Transactions on Multimedia (TMM)*, 24:3665–3678, 2022.

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 (Proc. IEEE)*, 112(11):1716–1754, 2025.

Yongxin Wang, Xin Luo, Liqiang Nie, Jingkuan Song, Wei Zhang, and Xin-Shun Xu. Batch: A scalable asymmetric discrete cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 33(11):3507–3519, 2021.

Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434, 2013.

Wai Keung Wong, Lunke Fei, Jianyang Qin, Shuping Zhao, Jie Wen, and Zhihao He. Heterogeneous pairwise-semantic enhancement hashing for large-scale cross-modal retrieval. *IEEE Transactions on Multimedia (TMM)*, 27:3238–3250, 2025.

Chao Zhang, Huaxiong Li, Yang Gao, and Chunlin Chen. Weakly-supervised enhanced semantic-aware hashing for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 35(6):6475–6488, 2023a.

Donglin Zhang, Xiao-Jun Wu, Tianyang Xu, and He-Feng Yin. Dah: Discrete asymmetric hashing for efficient cross-media retrieval. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 35(2):1365–1378, 2023b.

Donglin Zhang, Chang-Xing Li, Mengke Li, and Zhikai Hu. Discrete elective hashing with incomplete labels for efficient cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)*, 21(6), 2025.

Long Zheng, Hongmei Chen, Biao Xiang, Shi-Jinn Horng, and Tianrui Li. Multi-similarity preserving semantic-aware hashing for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(8):11593–11606, 2026.

Chang Zhou, Lai Man Po, and Weifeng Ou. Angular deep supervised vector quantization for image retrieval. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 33(4):1638–1649, 2020.

A Related Work

Cross-Modal Hashing and Semantic Alignment. Cross-modal hashing has been widely studied for efficient cross-modal retrieval (CMR) due to its low storage cost and fast Hamming-distance computation. Existing methods usually project heterogeneous modalities into a shared latent space and exploit semantic supervision to learn compact binary representations Tang et al. (2016); Chen et al. (2020); Wang et al. (2021). Recent works further improve retrieval performance by modeling label correlations, pairwise semantic relationships, distribution consistency, or regenerated hash codes Zhang et al. (2023b;a); Li et al. (2023); Sun et al. (2024b); Huang et al. (2024); Liu et al. (2024); Li et al. (2025b); Zheng et al. (2026). These methods demonstrate the importance of semantic alignment and relationship preservation for CMR. However, they mainly encode data into binary hash codes, whose representation capacity can be limited under short code lengths. In contrast, this paper focuses on product quantization-based CMR, which represents data by codewords and provides a more flexible storage-accuracy trade-off.

Product Quantization for Cross-Modal Retrieval. Product quantization (PQ) was originally proposed for efficient approximate nearest neighbor search by decomposing the feature space into multiple subspaces and approximating vectors with subspace codewords Jegou et al. (2010). It has been widely used in visual retrieval and extended to deep or residual quantization frameworks Gao et al. (2019); Song et al. (2019); Chen et al. (2022); Huijben et al. (2024). Compared with cross-modal hashing, however, PQ-based cross-modal retrieval remains relatively underexplored. Representative methods such as Composite Correlation Quantization, Collective Deep Quantization, and Asymmetric Correlation Quantization Hashing learn compact quantized representations by modeling cross-modal correlations Long et al. (2016); Cao et al. (2017); Wang et al. (2022). Nevertheless, existing PQ-based CMR methods mainly optimize codebooks for feature approximation, where codewords usually lack explicit semantic correspondence and semantic supervision is only indirectly imposed on the learned representations. Different from them, P2PQ constructs category-aware implicit codebooks whose size is determined by the number of semantic categories. The learned codebooks simultaneously serve as quantization prototypes and semantic regression weights, thereby coupling discrete quantization with semantic discrimination. Moreover, P2PQ preserves label-guided semantic relationship consistency between continuous representations and quantized codes without explicitly constructing an affinity graph.

B Optimization Details

B.1 Step 1:

Update $\mathbf{Z}_p$ with $\partial\mathcal{J}/\partial\mathbf{Z}_p = 0$, we have

$$\mathbf{Z}_p = (\mathbf{I} + \beta\mathbf{B}_2\mathbf{B}_2^\mathsf{T})^{-1}(\mathbf{B}_1 + \beta(\mathbf{B}_2\bar{\mathbf{L}})\bar{\mathbf{L}}^\mathsf{T})\mathbf{X}_p^\mathsf{T}(\mathbf{X}_p\mathbf{X}_p^\mathsf{T})^{-1}, \tag{26}$$

where $\mathbf{B}_1 = \mathbf{Q}_p\mathbf{H}\mathbf{S}_p\mathbf{V}^\mathsf{T} \in \mathbb{R}^{r\times N}$, $\mathbf{B}_2 = \mathbf{Q}_p\mathbf{S}_p\bar{\mathbf{V}}^\mathsf{T} \in \mathbb{R}^{r\times N}$, and $\mathbf{I} \in \mathbb{R}^{r\times r}$ is an identity matrix.

B.2 Step 2:

Update $\mathbf{Q}_p$ with others fixed, the subproblem is

$$\begin{aligned} \min_{\mathbf{Q}_p} \quad & \|\hat{\mathbf{X}}_p - \mathbf{Q}_p\mathbf{H}\mathbf{S}_p\mathbf{V}^\mathsf{T}\|_\mathrm{F}^2 + \beta\|\bar{\mathbf{L}}\bar{\mathbf{L}}^\mathsf{T} - \hat{\mathbf{X}}_p^\mathsf{T}\mathbf{Q}_p\mathbf{S}_p\bar{\mathbf{V}}^\mathsf{T}\|_\mathrm{F}^2, \\ & s.t. \quad \mathbf{Q}_p^\mathsf{T}\mathbf{Q}_p = \mathbf{I}_r, \end{aligned} \tag{27}$$

where $\hat{\mathbf{X}}_p = \mathbf{Z}_p\mathbf{X}_p$. Due to the orthogonality constraint, the subproblem is optimized on the Stiefel manifold using the method in Wen & Yin (2013). Specifically, denoting the value of $\mathbf{Q}_p$ after the ith iteration as $\mathbf{Q}_p(i)$, then $\mathbf{Q}_p(i+1)$ can be obtained by

$$\mathbf{Q}_p(i+1) = \left(\mathbf{I} + \frac{\kappa}{2}\boldsymbol{\nabla}\right)^{-1}\left(\mathbf{I} - \frac{\kappa}{2}\boldsymbol{\nabla}\right)\mathbf{Q}_p(i), \tag{28}$$

where κ is the step size determined by a nonmonotone line search, and $\boldsymbol{\nabla} = \mathbf{G}(\mathbf{Q}_p)\mathbf{Q}_p(i)^\mathsf{T} - \mathbf{Q}_p(i)\mathbf{G}(\mathbf{Q}_p)^\mathsf{T}$ is a skew-symmetric matrix, with

$$\mathbf{G}(\mathbf{Q}_p) = 2(\mathbf{B}_1 - \hat{\mathbf{X}}_p)\mathbf{V}\mathbf{S}_p\mathbf{H}^\mathsf{T} + 2\beta\hat{\mathbf{X}}_p(\hat{\mathbf{X}}_p^\mathsf{T}\mathbf{B}_2 - \bar{\mathbf{L}}\bar{\mathbf{L}}^\mathsf{T})\bar{\mathbf{V}}\mathbf{S}_p. \tag{29}$$

B.3 Step 3:

Update $\mathbf{H}$ with $\partial\mathcal{J}/\partial\mathbf{H} = 0$, we have

$$\mathbf{H} = \left(\sum_{p=1}^{P}\mathbf{Q}_p^\mathsf{T}\hat{\mathbf{X}}_p\mathbf{V}\mathbf{S}_p\right)\left(\sum_{p=1}^{P}\mathbf{S}_p\mathbf{V}^\mathsf{T}\mathbf{V}\mathbf{S}_p + \mu\mathbf{I}_r\right)^{-1}. \tag{30}$$

B.4 Step 4:

To isolate the terms associated with a specific block $\mathbf{V}_m$, we define a constant matrix $\mathbf{B}_p = \mathbf{S}_p\mathbf{H}^\mathsf{T}\mathbf{Q}_p^\mathsf{T}$. Corresponding to the block structure of $\mathbf{V}$, we vertically partition $\mathbf{B}_p$ into M row blocks,

denoted as $\mathbf{B}_p = [\mathbf{B}_{p,1}^\mathsf{T}, \mathbf{B}_{p,2}^\mathsf{T}, \ldots, \mathbf{B}_{p,M}^\mathsf{T}]^\mathsf{T}$. Consequently, the matrix multiplication can be equivalently expanded as a sum of block-wise products: $\mathbf{V}\mathbf{B}_p = \sum_{j=1}^{M} \mathbf{V}_j \mathbf{B}_{p,j}$. By separating the m-th block from the others, we obtain $\mathbf{V}\mathbf{B}_p = \mathbf{V}_m \mathbf{B}_{p,m} + \sum_{j \neq m} \mathbf{V}_j \mathbf{B}_{p,j}$. Substituting this expansion back into the original equation and omitting the terms independent of $\mathbf{V}_m$, the objective function for the local subproblem concerning $\mathbf{V}_m$ is reformulated as:

$$\begin{aligned} \min \mathcal{J}(\mathbf{V}_m) = & \sum_{p=1}^{P} \left\| \left(\hat{\mathbf{X}}_p^\mathsf{T} - \sum_{j \neq m} \mathbf{V}_j \mathbf{B}_{p,j} \right) - \mathbf{V}_m \mathbf{B}_{p,m} \right\|_\mathrm{F}^2 \\ & + \alpha \| \mathbf{V}_m \mathbf{W}_m + \mathbf{1}_N \mathbf{t}^\mathsf{T} - \mathbf{L} - \mathbf{C} \odot \mathbf{D} \|_\mathrm{F}^2. \end{aligned} \tag{31}$$

Update $\mathbf{V}_m$ with $\partial \mathcal{J}(\mathbf{V}_m)/\partial \mathbf{V}_m = 0$, we obtain

$$\begin{aligned} \mathbf{V}_m = & \left(\sum_{p=1}^{P} \left(\hat{\mathbf{X}}_p^\mathsf{T} - \sum_{j \neq m} \mathbf{V}_j \mathbf{B}_{p,j} \right) \mathbf{B}_{p,m}^\mathsf{T} \right. \\ & \left. -\alpha (\mathbf{1}_N \mathbf{t}^\mathsf{T} - \mathbf{Y}) \mathbf{W}_m^\mathsf{T} \right) \left(\sum_{p=1}^{P} \mathbf{B}_{p,m} \mathbf{B}_{p,m}^\mathsf{T} + \alpha \mathbf{W}_m \mathbf{W}_m^\mathsf{T} \right)^{-1}. \end{aligned} \tag{32}$$

where $\mathbf{Y} = \mathbf{L} + \mathbf{C} \odot \mathbf{D}$. Then, $\mathbf{V}$ is constructed by column-wise concatenation, i.e., $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_M]$.

B.5 Step 5:

Update $\mathbf{t}$ with $\partial \mathcal{J}/\partial \mathbf{t} = 0$, we have

$$\mathbf{t} = \frac{1}{MN} \sum_{m=1}^{M} (\mathbf{Y} - \mathbf{V}_m \mathbf{W}_m)^\mathsf{T} \mathbf{1}_N. \tag{33}$$

B.6 Step 6:

For updating $\mathbf{D}$, setting the partial derivative of $\mathcal{J}$ with respect to unconstrained D to zero, we obtain:

$$-2\alpha \sum_{m=1}^{M} \mathbf{C} \odot (\mathbf{V}_m \mathbf{W}_m + \mathbf{1}_N \mathbf{t}^\mathsf{T} - \mathbf{L} - \mathbf{C} \odot \mathbf{D}) = 0. \tag{34}$$

Since $\mathbf{C} \in \{-1, 1\}^{N \times c}$ is a sign matrix, it naturally satisfies the property $\mathbf{C} \odot \mathbf{C} = \mathbf{1}$. By expanding the equation and distributing $\mathbf{C}$ into the terms, the summation $\sum_{m=1}^{M} \mathbf{C} \odot \mathbf{C} \odot \mathbf{D}$ can be simplified to $M\mathbf{D}$. Additionally, since $\mathbf{D}$ must satisfy the non-negativity constraint ($\mathbf{D} \geq 0$), we project the unconstrained solution onto the feasible set using the $\max(\cdot, 0)$ operator. This yields the final closed-form update rule for $\mathbf{D}$:

$$\mathbf{D} = \max \left(\frac{1}{M} \sum_{m=1}^{M} \mathbf{C} \odot (\mathbf{V}_m \mathbf{W}_m + \mathbf{1}_N \mathbf{t}^\mathsf{T} - \mathbf{L}), \quad 0 \right). \tag{35}$$

B.7 Step 7:

Fixing other variables, the sub-problem of $\mathbf{S}_p = \mathrm{diag}(\mathbf{s}_p)$ is formulated as:

$$\begin{aligned} \min \mathcal{J}(\mathbf{S}_p) = & \| \mathbf{Z}_p \mathbf{X}_p - \mathbf{Q}_p \mathbf{H} \mathbf{S}_p \mathbf{V}^\mathsf{T} \|_\mathrm{F}^2 \\ & + \beta \| \bar{\mathbf{L}} \bar{\mathbf{L}}^\mathsf{T} - \mathbf{X}_p^\mathsf{T} \mathbf{Z}_p^\mathsf{T} \mathbf{Q}_p \mathbf{S}_p \bar{\mathbf{V}}^\mathsf{T} \|_\mathrm{F}^2. \end{aligned} \tag{36}$$

Direct optimization of a diagonal matrix is algebraically cumbersome. By utilizing the trace properties, we can rewrite Eq. (36) into a standard quadratic programming problem:

$$\min_{\mathbf{s}_p} \mathcal{J}(\mathbf{s}_p) = \mathbf{s}_p^\mathsf{T} \mathbf{\Gamma}_p \mathbf{s}_p - 2\mathbf{s}_p^\mathsf{T} \boldsymbol{\gamma}_p, \tag{37}$$

where $\mathbf{s}_p \in \mathbb{R}^{r\times 1}$. $\mathbf{\Gamma}_p$ and $\boldsymbol{\gamma}_p$ are defined as:

$$\begin{aligned}\mathbf{\Gamma}_p = &\left((\mathbf{H}^\top\mathbf{H}) \odot (\mathbf{V}^\top\mathbf{V})\right) \\ &+ \beta\left((\mathbf{Q}_p^\top\hat{\mathbf{X}}_p\hat{\mathbf{X}}_p^\top\mathbf{Q}_p) \odot (\bar{\mathbf{V}}^\top\bar{\mathbf{V}})\right),\end{aligned} \tag{38}$$

$$\boldsymbol{\gamma}_p = \text{diag}\left(\mathbf{H}^\top\mathbf{Q}_p^\top\hat{\mathbf{X}}_p\mathbf{V}\right) + \beta\text{diag}\left((\mathbf{Q}_p^\top\hat{\mathbf{X}}_p\bar{\mathbf{L}})(\bar{\mathbf{L}}^\top\bar{\mathbf{V}})\right). \tag{39}$$

Setting the derivative $\partial\mathcal{J}/\partial\mathbf{s}_p = 0$, we obtain the exact closed-form solution:

$$\mathbf{s}_p = \mathbf{\Gamma}_p^{-1}\boldsymbol{\gamma}_p \quad \Longrightarrow \quad \mathbf{S}_p = \text{diag}(\mathbf{s}_p). \tag{40}$$

B.8 STEP 8:

Updating $\mathbf{W}_m$. The subproblem is

$$\begin{aligned}\min_{\mathbf{W}_m} \mathcal{J}(\mathbf{W}_m) = \ &\alpha\|\mathbf{V}_m\mathbf{W}_m - \bar{\mathbf{Y}}\|_F^2 \\ &+ \beta\sum_{p=1}^{P}\|\mathbf{E}_{p,-m} - \mathbf{A}_{p,m}\mathbf{W}_m\mathbf{N}_m\|_F^2,\end{aligned} \tag{41}$$

where $\bar{\mathbf{Y}} = \mathbf{Y} - \mathbf{1}_N\mathbf{t}^\top$, $\mathbf{E}_{p,-m} = \bar{\mathbf{L}}\bar{\mathbf{L}}^\top - \sum_{j\neq m}\mathbf{A}_{p,j}\mathbf{W}_j\mathbf{N}_j$ and $\mathbf{A}_p = \hat{\mathbf{X}}_p^\top\mathbf{Q}_p\mathbf{S}_p = [\mathbf{A}_{p,1},\ldots,\mathbf{A}_{p,M}]$.

By setting the derivative of the objective function w.r.t. W_m to zero, we reorganize the terms into a classical Sylvester equation:

$$\mathbf{\Phi}_m\mathbf{W}_m + \mathbf{\Psi}_m\mathbf{W}_m\mathbf{\Lambda}_m = \mathbf{\Omega}_m, \tag{42}$$

where $\mathbf{\Phi}_m = \alpha\mathbf{V}_m^\top\mathbf{V}_m$, $\mathbf{\Psi}_m = \beta\sum_{p=1}^{P}\mathbf{A}_{p,m}^\top\mathbf{A}_{p,m}$, $\mathbf{\Lambda}_m = \mathbf{N}_m\mathbf{N}_m^\top$, and $\mathbf{\Omega}_m = \alpha\mathbf{V}_m^\top\bar{\mathbf{Y}} + \beta\sum_{p=1}^{P}\mathbf{A}_{p,m}^\top\mathbf{E}_{p,-m}\mathbf{N}_m^\top$.

By exploiting the algebraic property of the one-hot assignment matrix $\mathbf{N}_m$, its Gram matrix $\mathbf{\Lambda}_m$ is strictly a diagonal matrix $\text{diag}(c_1,\ldots,c_c)$, where c_i denotes the number of items assigned to the i-th cluster. This property completely decouples the equation column-by-column. The i-th column of $\mathbf{W}_m$ (denoted as $\mathbf{W}_{m,i}$) can be updated independently in parallel via an extremely fast closed-form:

$$\mathbf{W}_{m,i} = (\mathbf{\Phi}_m + c_i\mathbf{\Psi}_m + \epsilon\mathbf{I}_d)^{-1}\mathbf{\Omega}_{m,i}, \quad i = 1,\ldots,c, \tag{43}$$

where ϵ is a very small positive constant.

B.9 STEP 9:

Updating $\mathbf{N}_m$. The subproblem is

$$\begin{aligned}\min_{\mathbf{N}_m} \ &\sum_{p=1}^{P}\|\mathbf{E}_{p,-m} - \mathbf{A}_{p,m}\mathbf{W}_m\mathbf{N}_m\|_F^2 \\ &\text{s.t. } \mathbf{1}_c^\top\mathbf{N}_m = \mathbf{1}_N^\top.\end{aligned} \tag{44}$$

After expanding the Frobenius norm and removing constant terms, the problem reduces to

$$\min_{\mathbf{N}_m} \text{Tr}\left(\mathbf{\Delta}_m^\top\mathbf{N}_m\right), \tag{45}$$

where $\mathbf{\Delta}_m = \text{diag}(\mathbf{G}_m)\mathbf{1}_N^\top - 2\mathbf{R}_m$, where $\mathbf{G}_m = \sum_{p=1}^{P}\mathbf{W}_m^\top\mathbf{A}_{p,m}^\top\mathbf{A}_{p,m}\mathbf{W}_m$ and $\mathbf{R}_m = \sum_{p=1}^{P}\mathbf{W}_m^\top\mathbf{A}_{p,m}^\top\mathbf{E}_{p,-m}$. Thus, each sample can be assigned independently:

$$j_i^* = \arg\min_{j\in\{1,\ldots,c\}}(\mathbf{\Delta}_m)_{j,i}, \quad i = 1,\ldots,N. \tag{46}$$

The assignment matrix is then updated by setting $(\mathbf{N}_m)_{j_i^*,i} = 1$ and all remaining entries to zero.

We summarize the complete procedure of optimization in Algorithm 1.

Algorithm 1 The optimization of P2PQ

Input: Training data $\{X_p\}_{p=1}^{P}$, label matrix L
Parameter: Model parameters α, β, μ, and maximum iteration number τ
Output: Codebooks $\{W_m\}_{m=1}^{M}$, quantization codes $\{N_m\}_{m=1}^{M}$, and Mapping matrices $\{\mathbf{Z}_p\}_{p=1}^{P}$, $\{\mathbf{Q}_p\}_{p=1}^{P}$, and $\{\mathbf{S}_p\}_{p=1}^{P}$

Initialization: Randomly initialize S_p, V_m, H, t, D, W_m, and N_m, and randomly generate an orthogonal matrix Q_p;
Repeat
  Update $\{\mathbf{Z}_p\}_{p=1}^{P}$ via Eq. (26);
  Update $\{\mathbf{Q}_p\}_{p=1}^{P}$ via Eq. (28);
  Update $\{\mathbf{W}_m\}_{m=1}^{M}$ via Eq. (43);
  Update $\{\mathbf{N}_m\}_{m=1}^{M}$ via Eq. (46);
  Update $\{\mathbf{V}_m\}_{m=1}^{M}$ via Eq. (32);
  Update $\mathbf{H}$ via Eq. (30);
  Update $\{\mathbf{S}_p\}_{p=1}^{P}$ via Eq. (40);
  Update $\mathbf{t}$ via Eq. (33);
  Update $\mathbf{D}$ via Eq. (35);
Until up to τ

C Detailed Description of Datasets

MIRFLICKR-25K[1] contains 16,738 image-text pairs from 24 semantic categories, represented by 150-dimensional edge histogram image features and 500-dimensional textual features.
IAPRTC-12[2] contains 20,000 image-text pairs with 255 labels, represented by 512-dimensional GIST image features and 2912-dimensional BoW textual features.
NUS-WIDE[3] contains 186,577 image-text pairs from the 10 most frequent categories, represented by 500-dimensional SIFT image features and 1000-dimensional binary tag vectors.

Following common practice, 2,000 samples are randomly selected as queries for MIRFLICKR-25K and IAPRTC-12, and 1,867 samples are selected as queries for NUS-WIDE. The remaining samples are used for training and retrieval.

D Additional Experimental Results

D.1 More PR and P@N Curves

To provide a more comprehensive evaluation of the proposed method, we present additional retrieval results that are omitted from the main paper due to the page limit. Specifically, Fig. 5 reports the Precision–Recall (PR) curves under the 64-bit setting, complementing the 16-bit PR curves presented in the main paper. Moreover, Fig. 6 and Fig. 7 show the Precision@N (P@N) curves under the 16-bit and 64-bit settings, respectively, where N ranges from 1 to 2000. These additional results consistently demonstrate that P2PQ achieves superior retrieval performance across different code lengths, datasets, and retrieval tasks, further validating the effectiveness of the proposed framework.

D.2 Training Time Analysis

We further evaluate the efficiency of P2PQ on the large-scale NUS-WIDE dataset. As shown in Table 3, P2PQ requires more training time than most shallow hashing baselines due to its joint optimization of semantic alignment, codebook learning, and geometric preservation. However, it achieves consistently better retrieval accuracy and remains comparable to or faster than DCGH under longer code lengths. More importantly, the proposed category-aware codebook substantially

[1] https://press.liacs.nl/mirflickr/
[2] https://www.imageclef.org/photodata
[3] https://lms.comp.nus.edu.sg/research/NUS-WIDE.html

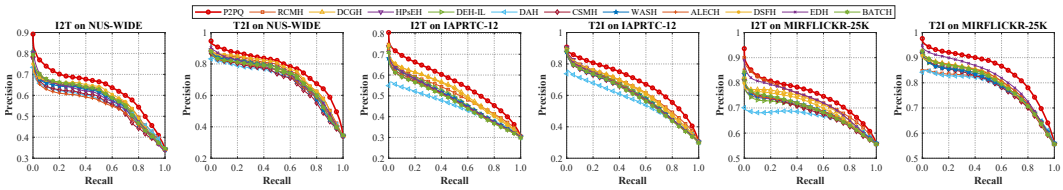


Figure 5: PR curves of P2PQ and the baseline methods on three datasets under 64-bit settings.

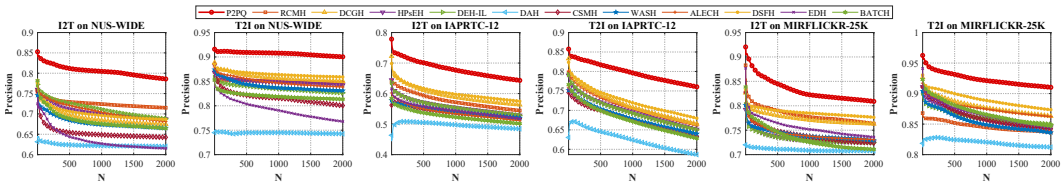


Figure 6: P@N curves of P2PQ and the baseline methods on three datasets under 16-bit settings.

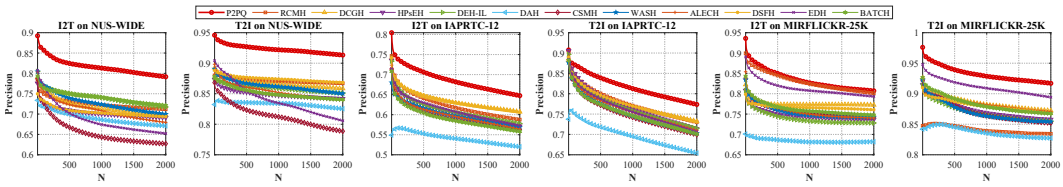


Figure 7: P@N curves of P2PQ and the baseline methods on three datasets under 64-bit settings.

reduces the number of codewords compared with conventional fixed-size PQ codebooks. These results demonstrate that P2PQ achieves a favorable trade-off between retrieval accuracy, codebook compactness, and training efficiency.

Table 3: Training time (seconds) on the NUS-WIDE dataset under different code lengths.

Method	**16 bits**	**32 bits**	**64 bits**	**128 bits**
BATCH	1.65	2.20	3.31	5.56
DAH	1.10	2.06	4.18	8.86
EDH	10.96	15.71	27.30	56.78
WASH	8.74	9.89	12.47	17.39
DCGH	4.69	24.55	134.92	528.77
ALECH	**0.84**	**1.45**	**2.83**	**5.41**
DSFH	4.65	6.94	12.85	24.64
RCMH	6.46	7.03	8.37	10.83
P2PQ	23.53	65.21	90.33	491.13

D.3 Comparison with Deep-Feature Methods

To further examine the compatibility of P2PQ with deep visual features, we conduct additional experiments on MIRFLICKR-25K using 4096-dimensional VGG image features and 1386-dimensional BoW textual features. The compared methods include three deep cross-modal hashing methods, i.e., DHaPH Huo et al. (2024), LCDH Li et al. (2025a), and DUaPH Han et al. (2026), as well as two shallow hashing methods using deep visual features, i.e., RCMH Jiang et al. (2026) and MPSH Zheng et al. (2026). As shown in Table 4, P2PQ achieves the best performance under all code lengths for both I2T and T2I retrieval. These results indicate that although P2PQ is a shallow product quantization framework, it can effectively benefit from stronger visual features and learn discriminative compact representations.

Table 4: mAP comparison on the MIRFLICKR-25K dataset under different code lengths.

Task	Method	MIRFLICKR-25K		
		16	32	64
I2T	DHaPH	0.8299	0.8437	0.8531
	LCDH	0.7898	0.7992	0.8034
	DUaPH	0.8359	0.8516	0.8589
	RCMH	0.8424	0.8554	0.8695
	MPSH	0.8533	0.8645	0.8699
	P2PQ	**0.8775**	**0.8838**	**0.8836**
T2I	DHaPH	0.8184	0.8165	0.8229
	LCDH	0.8223	0.8480	0.8556
	DUaPH	0.8163	0.8304	0.8357
	RCMH	0.8145	0.8281	0.8436
	MPSH	0.8236	0.8344	0.8384
	P2PQ	**0.8665**	**0.8725**	**0.8719**

E LIMITATIONS AND FUTURE WORK

Despite the promising performance of P2PQ, several limitations remain and provide directions for future research.

- **Dependence on pre-extracted features.** As a shallow framework, P2PQ relies on pre-extracted features and RBF kernel mapping, and its retrieval performance is therefore partly bounded by the quality of the input representations. In the future, we plan to integrate P2PQ with pretrained multimodal encoders and develop an end-to-end framework that jointly optimizes feature extraction, semantic alignment, and product quantization.
- **Category-dependent codebook design.** The number of codewords in each subspace is determined by the number of semantic categories. Although this design produces compact and semantically discriminative codebooks for datasets with moderate label spaces, its storage and assignment costs may increase when the category vocabulary becomes extremely large or dynamically evolves. Moreover, the current formulation assumes sufficiently reliable annotations. Future work will investigate adaptive, hierarchical, or factorized codebooks that decouple codebook capacity from the category number, together with robust learning strategies for incomplete, noisy, and open-set labels.
- **Training efficiency and computational implementation.** Although the implicit affinity computation avoids explicitly constructing an $N \times N$ graph and makes the overall complexity approximately linear in the number of training samples, the full-batch alternating optimization still involves repeated matrix operations, decompositions, and discrete assignments, resulting in relatively high training costs, particularly under longer code lengths. We will explore stochastic, incremental, and online optimization algorithms for large-scale and streaming data. Meanwhile, the modality-wise and subspace-wise updates provide opportunities for parallel execution. High-performance computing and heterogeneous CPU–GPU implementations, including parallel matrix kernels, mixed-precision computation, and distributed optimization, will therefore be investigated to further improve training efficiency and scalability.
- **Generality of the evaluated scenarios.** The current experiments mainly focus on paired image–text retrieval over three benchmark datasets. Future studies will extend P2PQ to additional modalities, incomplete or unpaired multimodal data, and substantially larger real-world datasets to more comprehensively examine its generalization and practical applicability.