SG-MoE: Shared Multi-Granularity MoE with MLLM-Assisted Semantic Router Learning for Cross-Modal Representation Learning and Zero-Shot Retrieval

Anonymous authors
Paper under double-blind review

Abstract

Zero-shot cross-modal retrieval matches semantically related samples across modalities under disjoint training and test categories, yet existing methods face insufficient multi-granularity coordination, limited cross-modal alignment beyond output representations, and semantically ambiguous expert routing. To address these issues, we propose shared multi-granularity MoE with MLLM-assisted semantic router learning for cross-modal representation learning and zero-shot retrieval (SG-MoE). We design a dual-stream MoE to learn fine-grained matching and semantic-level representations; introduce modality-shared experts to constrain internal cross-modal transformations while retaining modality-specific routers; and develop MLLM-assisted semantic router learning with an offline MLLM-built ontology and frozen-encoder supervision during training. Across four benchmarks, SG-MoE improves over task-specific baselines; ablations identify dual-stream modeling and modality-shared experts as the main retrieval contributors, while routing analyses support concept–expert organization rather than substantial Avg mAP gains. Retrieval uses the trained SG-MoE routers with fixed zero semantic input and no external MLLM or vision–language encoder.

1 Introduction

Zero-shot cross-modal retrieval (ZS-CMR) matches semantically related samples across heterogeneous modalities when training and test categories are disjoint (Lin et al., 2020; Ge et al., 2025). Unlike closed-category retrieval, ZS-CMR must generalize to objects, scenes, and semantic compositions absent from training (Lin et al., 2020; Ge et al., 2025). The resulting shift can alter cross-modal correspondences and the roles of fine-grained and semantic-level evidence, demanding transferable representations and coordinated multi-granularity learning.

Recent methods learn shared or disentangled representations for fine-grained, shifted, and partially paired retrieval (Wang et al., 2024; Ge et al., 2025; Li et al., 2025; Qin et al., 2026); fine-grained approaches exploit local–global, lexicon–patch, and granularity-aware correspondences (Gao et al., 2024; Yiyuan et al., 2024; Liu et al., 2026), while multimodal MoEs coordinate specialized experts through input-dependent routing (Li et al., 2024; Huai et al., 2025; Jing et al., 2026). Despite these advances, three interrelated limitations remain. **(i) Insufficient coordination of multi-granularity representations within conditional computation.** Existing methods typically encode fine-grained and semantic-level evidence through separate fixed paths before combining their scores (Gao et al., 2024; Yiyuan et al., 2024; Ge et al., 2025; Liu et al., 2026). They therefore lack unified conditional computation that coordinates locally discriminative and transferable semantic representations during learning, limiting multi-granularity modeling under unseen-category shifts. **(ii) Limited cross-modal semantic alignment beyond the output space.** Existing methods primarily align final image and text representations (Wang et al., 2024; Gao et al., 2024; Yiyuan et al., 2024; Liu et al., 2026). Matched samples may still pass through separate modality-specific internal functions, leaving their transformations unconstrained and weakening cross-modal transfer under zero-shot evaluation. **(iii) Semantically ambiguous expert routing under zero-shot shifts.** Although routing can exhibit semantic structure (Lee et al., 2025; Olson et al., 2025; Park & Park, 2026), routers

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.

optimized mainly by seen-category retrieval supervision need not form stable concept–expert associations. Their activation patterns may become ambiguous for novel categories, limiting specialization and interpretability. An offline concept ontology and training distributions from a frozen vision–language encoder can structure router learning while excluding external models from inference.

To address these challenges, we propose shared multi-granularity MoE with MLLM-assisted semantic router learning for cross-modal representation learning and zero-shot retrieval (SG-MoE). SG-MoE consists of three components. First, dual-stream multi-granularity MoE for cross-modal representation learning and zero-shot retrieval learns fine-grained matching representations and transferable semantic-level cross-modal retrieval representations within conditional computation, preserving local discrimination and semantic transfer under zero-shot shifts. Second, cross-modal representation alignment based on modality-shared experts shares expert parameters across image and text processing, extending alignment from final representations to shared internal expert functions while preserving modality-specific expert-routing flexibility. Third, MLLM-assisted semantic router learning for cross-modal representations uses an offline MLLM-constructed concept ontology and vision–language-encoder-generated concept distributions to condition and supervise router learning; retrieval inference uses the trained SG-MoE routers with a fixed zero semantic input.

Experiments on four zero-shot cross-modal retrieval benchmarks show that SG-MoE consistently improves over task-specific baselines. Component ablations associate the main retrieval gains with the multi-granularity MoE structure and modality-shared experts; semantic router learning yields limited Avg mAP change, while routing analyses support concept-related expert organization and interpretability.

The main contributions are summarized as follows:

- We propose shared multi-granularity MoE with MLLM-assisted semantic router learning for cross-modal representation learning and zero-shot retrieval (SG-MoE). It coordinates multi-granularity representations to learn transferable cross-modal correspondences.
- We design dual-stream multi-granularity MoE for cross-modal representation learning and zero-shot retrieval. It coordinates fine-grained and semantic-level retrieval representations through conditional computation.
- We establish cross-modal representation alignment based on modality-shared experts. Shared expert parameters constrain internal cross-modal transformations, while modality-specific routers accommodate heterogeneous inputs.
- We design MLLM-assisted semantic router learning for cross-modal representations. Two MLLMs construct and cross-review the offline concept ontology, and a frozen vision–language encoder generates training concept distributions for semantic conditioning and auxiliary router supervision. Inference uses the learned routers with a fixed zero semantic input and no external model.

2 Related Work

Zero-Shot Cross-Modal Retrieval. Existing methods learn shared embeddings through metric-learning, hashing, or adversarial objectives (Zhen et al., 2019; Liu & Wu, 2020; Lin et al., 2020), while recent studies examine semi-paired retrieval (Qin et al., 2026) and query-shift adaptation (Li et al., 2025). Fine-grained approaches such as VSE++, FILIP, PAN, and FGAN strengthen local visual–language correspondence (Faghri et al., 2018; Yao et al., 2022; Pan et al., 2023; Ge et al., 2025); FineLIP further uses dynamic local-token aggregation and token-level alignment for detailed captions and zero-shot retrieval (Asokan et al., 2025). These methods improve matching but generally process fine-grained and semantic-level evidence through separate fixed paths. SG-MoE instead coordinates both within a unified conditional-computation architecture under unseen-category shifts.

Multimodal Mixture-of-Experts. Multimodal MoEs study input-dependent routing and expert specialization for scalable modeling, continual learning, and modality-aware adaptation (Li et al., 2024; Jing et al., 2026; Huai et al., 2025; Wang et al., 2025; Yu et al., 2024; Chen et al., 2024). LIMoE processes image and text tokens with a shared sparse backbone and analyzes modality-specialized experts (Mustafa et al., 2022). VLMo combines modality-specific feed-forward experts

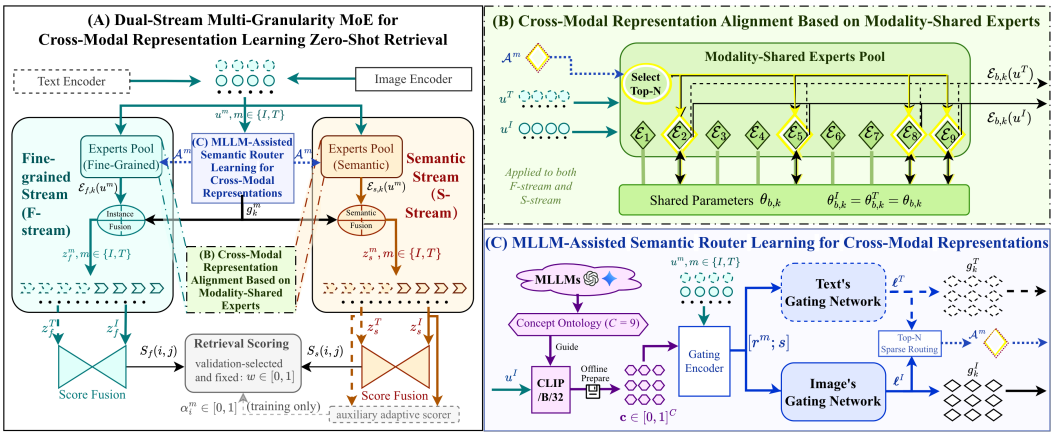


Figure 1: Overall framework of SG-MoE: (A) dual-stream multi-granularity MoE for cross-modal representation learning and zero-shot retrieval, (B) cross-modal representation alignment based on modality-shared experts, and (C) MLLM-assisted semantic router learning for cross-modal representations.

with shared self-attention and supports dual-encoder image–text retrieval (Bao et al., 2022). These works demonstrate shared sparse modeling and partial cross-modal parameter sharing; SG-MoE specifically shares corresponding expert transformations across modalities within each retrieval-granularity stream while retaining modality-specific routers.

Semantic Router Learning. Semantic guidance can derive from attributes, embeddings, concepts, or language-model supervision. Prior work studies semantic expert specialization (Park & Park, 2026), routing over expert embedding models (Lee et al., 2025), and routing diagnostics (Olson et al., 2025). Stable concept–expert associations for novel categories remain underexplored. SG-MoE instead uses an MLLM-assisted offline ontology and frozen-encoder concept distributions to organize and supervise router training under unseen-category shifts, without external semantic models at inference.

3 Methodology

As illustrated in Figure 1, Shared Multi-Granularity MoE with MLLM-Assisted Semantic Router Learning for Cross-Modal Representation Learning and Zero-Shot Retrieval (SG-MoE) comprises three modules. Dual-Stream Multi-Granularity MoE for Cross-Modal Representation Learning and Zero-Shot Retrieval learns locally discriminative fine-grained representations and transferable semantic-level representations within conditional computation for unseen-category retrieval. Cross-Modal Representation Alignment Based on Modality-Shared Experts shares expert transformations across modalities while preserving modality-specific routing. MLLM-Assisted Semantic Router Learning for Cross-Modal Representations uses an offline MLLM-assisted concept ontology and vision–language-encoder-generated concept distributions for router training.

3.1 Dual-Stream Multi-Granularity MoE for Cross-Modal Representation Learning and Zero-Shot Retrieval

For an input from modality $m \in \{I, T\}$, with I and T denoting image and text, let u^m be its sequence representation. SG-MoE comprises a Fine-Grained Stream (F-stream) and a Semantic Stream (S-stream): each F-stream expert applies self-attention, element scoring, and differentiable top-k pooling to retain the highest-scoring local elements, whereas each S-stream expert aggregates the complete sequence through a learnable query. The F-stream therefore emphasizes local instance-discriminative cues, while the S-stream captures global semantic abstractions intended to transfer

across unseen categories. For stream $b \in \{f, s\}$, the MoE output z_b^m is:

$$z_b^m = \sum_{k=1}^{K} g_k^m \mathcal{E}_{b,k}(u^m; \theta_{b,k}), \tag{1}$$

where $K = 9$ is the number of experts, $\mathcal{E}_{b,k}$ is expert k in stream b with parameters $\theta_{b,k}$, and g_k^m is its sparse routing weight. The choices $b = f$ and $b = s$ produce z_f^m and z_s^m, respectively. Both streams share g_k^m for the same input but retain separate expert sets, coordinating their expert indices while preserving granularity-specific transformations; image and text remain governed by modality-specific routers. The S-stream remains a learned retrieval representation, and offline concepts only organize router learning during training.

After obtaining the two representations, the stream scores $S_f(i, j)$ and $S_s(i, j)$ are:

$$\begin{aligned} S_f(i, j) &= \frac{(z_{f,i}^I)^\top z_{f,j}^T}{\|z_{f,i}^I\|_2 \|z_{f,j}^T\|_2}, \\ S_s(i, j) &= -\|z_{s,i}^I - z_{s,j}^T\|_2. \end{aligned} \tag{2}$$

The fixed full-gallery retrieval score is:

$$\mathrm{Score}_w(i, j) = w\widetilde{S}_f(i, j) + (1 - w)\widetilde{S}_s(i, j), \tag{3}$$

where S_f is cosine similarity, S_s is negative Euclidean distance, and $\widetilde{S}_f, \widetilde{S}_s$ are obtained by label-free z-score normalization over each evaluated score block. The weight $w \in [0, 1]$ is selected by Avg mAP on the seen-category validation subset and fixed before final evaluation.

3.2 Cross-Modal Representation Alignment Based on Modality-Shared Experts

Within each stream b, SG-MoE shares a single parameter set $\theta_{b,k}$ for expert k across image and text, while the F-stream and S-stream retain separate expert sets and the routers remain modality-specific. The image- and text-path expert parameters $\theta_{b,k}^I$ and $\theta_{b,k}^T$ satisfy:

$$\theta_{b,k}^I = \theta_{b,k}^T = \theta_{b,k}. \tag{4}$$

Under this constraint, Eq. (1) covers both z_b^I and z_b^T, which are computed from (g_k^I, u^I) and (g_k^T, u^T), respectively, while invoking the same $\mathcal{E}_{b,k}(\cdot; \theta_{b,k})$. Matched cross-modal samples may select distinct expert subsets: overlapping expert indices invoke the same shared transformations, while different active sets preserve modality-dependent routing flexibility.

Modality-specific routing accommodates heterogeneous input statistics, while hard parameter tying establishes common internal transformation paths and extends alignment beyond output representations. Because both modality paths invoke the same $\theta_{b,k}$, image- and text-path gradients accumulate on the shared expert parameters, exposing each expert to cross-modal retrieval supervision while retaining the output-space objective.

3.3 MLLM-Assisted Semantic Router Learning for Cross-Modal Representations

SG-MoE organizes training-time routing in a fixed $C = 9$ concept space constructed once offline with two MLLMs. Given the same instruction to organize general entities into ten broad concepts and provide definitions, the two MLLMs independently generated candidate ontologies and reviewed each other's proposals. Because *Person* and *Clothing* overlap semantically at the intended broad granularity, they were consolidated into *Person & Wearables*. The resulting ontology comprises Person & Wearables, Animal, Plant, Food & Drink, Vehicle / Transport, Building & Structure, Furniture & Large Fixture, Tool & Small Artifact, and Natural Environment. The ontology is defined at a broad semantic level independently of dataset class names. Setting $K = C = 9$ aligns expert and concept indices. The MLLMs are used only for offline ontology construction and cross-review, not for sample-level expert selection.

For a training image, a frozen pretrained vision–language encoder scores the nine fixed concept prompts and produces the concept distribution $\mathbf{c} = [c_1, \ldots, c_C] \in [0,1]^C$ through softmax, where $\sum_{c=1}^{C} c_c = 1$. During training, the image-derived $\mathbf{c}$ is used by both routers of the paired image–text sample as privileged semantic-conditioning input and as auxiliary supervision through $\mathcal{L}_{gate}$. These training targets are generated by the frozen vision–language encoder rather than by the MLLMs.

For modality m, the router feature r^m, semantic-conditioning input $\mathbf{s}$, and expert logits $\boldsymbol{\ell}^m$ are computed as:

$$r^m = \mathrm{MLP}([\mu(u^m); \sigma(u^m)]), \qquad \mathbf{s} = \begin{cases} \mathbf{c}, & \text{training,} \\ \mathbf{0}, & \text{inference.} \end{cases} \qquad \boldsymbol{\ell}^m = \mathcal{G}^m([r^m; \mathbf{s}]), \tag{5}$$

where $[\,;\,]$ denotes concatenation, and $\mu(\cdot)$ and $\sigma(\cdot)$ are the sequence-wise mean and standard deviation. The MLP maps these statistics to r^m, and each modality-specific router $\mathcal{G}^m$ predicts the K expert logits. During training, $\mathbf{c}$ conditions both routers and supervises $\boldsymbol{\ell}^m$ through $\mathcal{L}_{gate}$ to organize concept–expert associations without directly determining expert selection. At inference, a fixed zero vector preserves the input dimensionality, and the learned routers generate sparse weights without the MLLMs or frozen vision–language encoder. All reported results follow this protocol.

Given $\boldsymbol{\ell}^m$, let $\mathcal{A}^m$ be the indices of its top-N logits, so $|\mathcal{A}^m| = N$. The sparse routing weight is:

$$g_k^m = \begin{cases} \dfrac{\exp(\ell_k^m)}{\sum_{k' \in \mathcal{A}^m} \exp(\ell_{k'}^m)}, & k \in \mathcal{A}^m, \\ 0, & k \notin \mathcal{A}^m. \end{cases} \tag{6}$$

The same budget N is assigned to both modalities, although their selected sets may differ; scoring is computed from the resulting representations.

Training objective. After curriculum warm-up, all three modules are jointly optimized by the objective below; the warm-up stage optimizes only the representative F- and S-stream retrieval paths. Let $C_b(i,j)$ denote the cosine similarity between the image and text outputs of stream b, and let $\bar{z}_{s,i}^m = K^{-1} \sum_{k=1}^{K} \mathcal{E}_{s,k}(u_i^m)$ be the mean S-stream expert feature for sample i and modality m. A shared head h predicts the training-time fusion weight $\alpha_i^m = h(\bar{z}_{s,i}^m) \in [0,1]$. The directional score matrices $P^{I\to T}$ and $P^{T\to I}$ are:

$$\begin{aligned} P_{ij}^{I\to T} &= \alpha_i^I C_f(i,j) + (1-\alpha_i^I) C_s(i,j), \\ P_{ij}^{T\to I} &= \alpha_i^T C_f(j,i) + (1-\alpha_i^T) C_s(j,i), \end{aligned} \tag{7}$$

where B denotes the number of matched image–text pairs in the current training mini-batch, and $P^{I\to T}, P^{T\to I} \in \mathbb{R}^{B\times B}$. In $P^{T\to I}$, row index i denotes the textual query and column index j denotes the image candidate. The query-conditioned cosine fusion jointly optimizes both streams during training. The shared prediction head h and its query-conditioned weights α_i^m are not used for full-gallery retrieval. Instead, retrieval applies Eq. (3): the z-score-normalized F-stream cosine similarity and S-stream negative Euclidean distance are combined using the fixed weight w selected on the seen-category validation subset. The primary retrieval loss $\mathcal{L}_{ret}$ is:

$$\mathcal{L}_{ret} = \tfrac{1}{2}\left[\mathrm{CE}(P^{I\to T}/\tau, y) + \mathrm{CE}(P^{T\to I}/\tau, y)\right], \tag{8}$$

where CE is row-wise cross-entropy over the B candidates, $y_i = i$ identifies the matched candidate, and $\tau > 0$ is the temperature. The router-supervision loss $\mathcal{L}_{gate}$ is:

$$\mathcal{L}_{gate} = \frac{1}{2B} \sum_{i=1}^{B} \sum_{m\in\{I,T\}} \mathrm{BCE}(\mathrm{sigmoid}(\boldsymbol{\ell}_i^m), \mathbf{c}_i), \tag{9}$$

where BCE is the mean element-wise binary cross-entropy over the $K = C$ aligned concept/expert dimensions, and the sigmoid function is applied element-wise to the router logits $\boldsymbol{\ell}_i^m$. The targets $\mathbf{c}_i$ are generated by the frozen vision–language encoder for the training images.

The expert-level objective $\mathcal{L}_{exp}$ is:

$$\mathcal{L}_{exp} = \frac{1}{2B} \sum_{i=1}^{B} \sum_{k=1}^{K} \bar{g}_{i,k} \left(\rho_{i,k}^{I\to T} + \rho_{i,k}^{T\to I}\right), \tag{10}$$

where $\bar{g}_{i,k} = (g^I_{i,k} + g^T_{i,k})/2$, and $\rho^{I\to T}_{i,k}$ and $\rho^{T\to I}_{i,k}$ are the directional per-sample retrieval losses of the k-th F-stream expert; the S-stream experts are optimized through the retrieval objective and auxiliary regularization under the same semantically supervised routing weights. The complete training objective $\mathcal{L}$ is:

$$\mathcal{L} = \mathcal{L}_{ret} + \lambda_{gate}\mathcal{L}_{gate} + \lambda_{exp}\mathcal{L}_{exp} + \lambda_{aux}\mathcal{L}_{aux}, \tag{11}$$

where λ_{gate}, λ_{exp}, and λ_{aux} are the weights of $\mathcal{L}_{gate}$, $\mathcal{L}_{exp}$, and $\mathcal{L}_{aux}$, respectively, with $\lambda_{aux} = 1$ fixed in all experiments. The implementation decomposes the auxiliary term as:

$$\mathcal{L}_{aux} = \beta_{sem}\mathcal{L}_{sem} + \beta_{bal}\mathcal{L}_{bal} + \beta_{ent}\mathcal{L}_{ent}, \tag{12}$$

where the β coefficients are the corresponding implementation weights. For $p^m_i = \mathrm{softmax}(z^m_{s,i})$ and the distribution q_i predicted from image i by the frozen ImageNet-pretrained ResNet-50 teacher, the S-stream semantic-consistency term is:

$$\begin{aligned}\mathcal{L}_{sem} = \frac{1}{2.5B}\sum_{i=1}^{B}\big[&D_{\mathrm{KL}}(q_i\|p^I_i) + D_{\mathrm{KL}}(q_i\|p^T_i)\\ &+ 0.25D_{\mathrm{KL}}(\mathrm{sg}[p^I_i]\|p^T_i) + 0.25D_{\mathrm{KL}}(\mathrm{sg}[p^T_i]\|p^I_i)\big],\end{aligned} \tag{13}$$

where sg denotes stop-gradient. This frozen image teacher is distinct from the frozen CLIP-B/32 encoder that produces the router concept targets $\mathbf{c}_i$. Let $\bar{g}^m_k = B^{-1}\sum_{i=1}^{B} g^m_{i,k}$. The remaining auxiliary terms are:

$$\mathcal{L}_{bal} = \frac{K}{2}\sum_{m\in\{I,T\}}\sum_{k=1}^{K}(\bar{g}^m_k)^2, \tag{14a}$$

$$\mathcal{L}_{ent} = \frac{1}{2B}\sum_{m\in\{I,T\}}\sum_{i=1}^{B}\sum_{k=1}^{K} g^m_{i,k}\log(g^m_{i,k}+\epsilon), \tag{14b}$$

where $\epsilon > 0$ is a numerical stabilizer. Thus, $\mathcal{L}_{sem}$ regularizes the S-stream representations, while $\mathcal{L}_{bal}$ and $\mathcal{L}_{ent}$ regularize expert utilization and routing weights. These training-only terms do not change the fixed full-gallery retrieval rule in Eq. (3).

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate SG-MoE on four zero-shot cross-modal retrieval benchmarks: Pascal Sentence Rashtchian et al. (2010), NUS-WIDE Chua et al. (2009), MS-COCO Lin et al. (2014), and MIRFlickr Huiskes & Lew (2008). Following the specialized ZS-CMR protocol in prior work Lin et al. (2020); Ge et al. (2025), paired samples are divided into training and validation partitions with the same label-level seen/unseen split in both. Model optimization is conducted on the seen-category training subset; checkpoints, hyperparameters, and fusion weights are selected on the seen-category validation subset. Unseen-category samples are reserved for final query/gallery evaluation. Pascal Sentence, NUS-WIDE, and MIRFlickr adopt 1:1 seen/unseen class splits with ten, three, and five repeated random splits, respectively; MS-COCO adopts a fixed $60/20$ seen/unseen split. We report image-to-text retrieval mAP (I2T), text-to-image retrieval mAP (T2I), and their average (Avg).

Baselines. Independently trained ZS-CMR baselines include DSCMR Zhen et al. (2019), JDSH Liu & Wu (2020), and FGAN Ge et al. (2025). The official TCR implementation Li et al. (2025), originally designed for query-shift adaptation, is re-run on the unseen-query and gallery embeddings from each dataset-specific SG-MoE encoder. This test-time adaptation evaluation, denoted SG-MoE+TCR, adds no trainable parameters. CLIP Radford et al. (2021) and SigLIP Zhai et al. (2023) are included as foundation-model references. The non-foundation comparisons use the same zero-shot splits, ImageNet-pretrained VGG19-BN Simonyan & Zisserman (2015) backbone with Block-5 fine-tuning, Word2Vec Mikolov et al. (2013) text features, retrieval protocol, and mAP computation as SG-MoE. CLIP and SigLIP provide context under their large-scale pretraining and backbone settings.

Table 1: Zero-shot cross-modal retrieval mAP on four benchmarks. Repeated-split results are averaged. Bold marks SG-MoE, and underlining marks the strongest non-foundation result; ties are also bolded.

Method	Pascal Sentence			NUS-WIDE			MS-COCO			MIRFlickr		
	I2T	T2I	Avg	I2T	T2I	Avg	I2T	T2I	Avg	I2T	T2I	Avg
CLIP	0.5648	0.5799	0.5724	0.6343	0.5890	0.6117	0.4810	0.4856	0.4833	0.4993	0.3797	0.4395
SigLIP	0.4113	0.5052	0.4582	0.5645	0.5675	0.5660	0.3958	0.4481	0.4219	0.4505	0.3989	0.4247
DSCMR	0.1235	0.1191	0.1213	0.4153	0.4087	0.4120	0.0855	0.0660	0.0758	0.1910	0.1437	0.1673
JDSH	0.1139	0.1215	0.1177	0.4252	0.3904	0.4078	0.0679	0.0680	0.0679	0.1616	0.1352	0.1484
SG-MoE+TCR	0.3477	0.2734	0.3105	**0.6067**	0.5559	0.5813	0.4160	0.4140	0.4150	0.2508	0.2203	0.2356
FGAN	0.3885	0.3951	0.3918	0.4895	0.4527	0.4711	0.4560	0.4771	0.4668	0.1770	0.1539	0.1655
SG-MoE	**0.4337**	**0.4343**	**0.4340**	**0.6067**	**0.5987**	**0.6027**	**0.4956**	**0.5186**	**0.5072**	**0.2963**	**0.2369**	**0.2666**

Table 2: SG-MoE settings by dataset. Batch Size: training/evaluation; Epochs: total/curriculum; RW Dropout: routing-weight dropout.

Setting	Pascal	NUS	COCO	MIR
Batch Size	$16/64$	$256/128$	$256/128$	$64/64$
Learning Rate	1×10^{-4}	1×10^{-5}	5×10^{-5}	8×10^{-5}
Warm-up Steps	200	600	600	900
Epochs	$20/6$	$18/2$	$15/6$	$30/4$
τ	0.18	0.18	0.25	0.18
λ_{exp}	0.5	0.3	0.3	0.4
Top-k Retention	10%	5%	2%	5%
Weight Decay	0.01	0.20	0.01	0.01
RW Dropout	0.2	0.1	0.1	0.1

Table 3: Cross-modal retrieval robustness across repeated unseen-category splits, reported as mean $\pm$ standard deviation of Avg mAP.

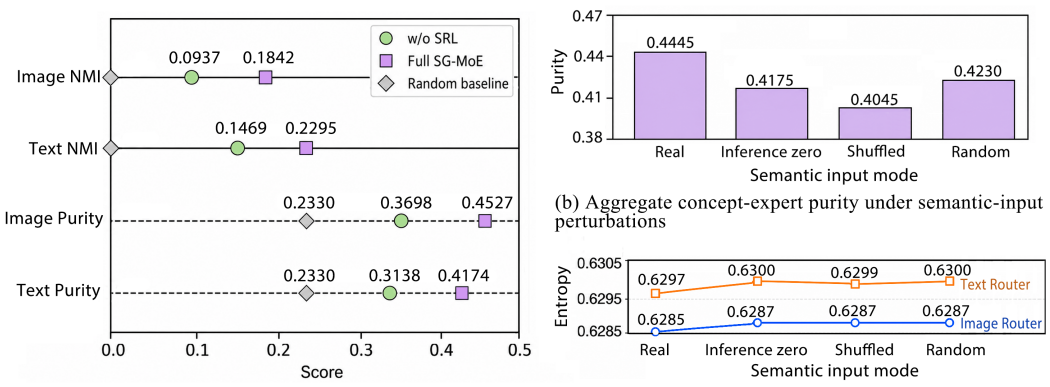
Method	NUS-WIDE	Pascal Sentence
CLIP	0.6117 ± 0.0478	0.5724 ± 0.0628
SigLIP	0.5660 ± 0.0769	0.4582 ± 0.0907
DSCMR	0.4120 ± 0.0440	0.1213 ± 0.0067
JDSH	0.4078 ± 0.0384	0.1177 ± 0.0041
SG-MoE+TCR	0.5813 ± 0.0468	0.3105 ± 0.0541
FGAN	0.4711 ± 0.1055	0.3918 ± 0.0450
SG-MoE	0.6027 ± 0.0560	0.4340 ± 0.0584

Implementation details. SG-MoE comprises $K = C = 9$ experts per stream with top-$N = 4$ sparse routing. All experiments are conducted on a single NVIDIA RTX 5090 GPU with 32GB memory. The offline ontology is constructed and cross-reviewed with GPT-5 and Gemini 2.5, while the frozen CLIP-B/32 encoder Radford et al. (2021) generates the training-time concept distributions used for semantic conditioning and router supervision. Table 2 lists the dataset-specific F-stream top-k retention proportions. Optimization employs AdamW, dataset-specific weight decay, and a cosine learning-rate schedule with linear warm-up; Warm-up Steps reports the corresponding optimizer-step count. Curriculum warm-up trains representative F- and S-stream paths whose parameters initialize the remaining experts before activating full top-N sparse aggregation and semantic-router supervision; λ_{gate} is zero during warm-up, is activated at 0.6 when full MoE training begins, and subsequently decays to 0.2. All dataset-specific hyperparameters, checkpoints, and fusion weights $w \in \{0, 0.1, \ldots, 1\}$ are selected on the seen-category validation subset and fixed before final unseen-category evaluation for every repeated split and ablation variant. The router parameters are learned jointly with SG-MoE. At retrieval inference, neither external model is used, and the routers receive a fixed zero semantic input.

4.2 Cross-Modal Retrieval Performance

Table 1 reports zero-shot retrieval performance on four benchmarks. SG-MoE exceeds every independently trained task-specific baseline across all four datasets. Relative to SG-MoE+TCR, SG-MoE improves both retrieval directions on Pascal Sentence, MS-COCO, and MIRFlickr; on NUS-WIDE, it matches I2T at 0.6067 and raises T2I from 0.5559 to 0.5987 (7.7%). Thus, the evaluated query-shift adaptation does not improve Avg mAP on any dataset, and SG-MoE requires no test-time adaptation. Among the foundation-model references, SG-MoE exceeds CLIP and SigLIP on MS-COCO and approaches CLIP on NUS-WIDE; these results provide context under different large-scale pretraining and backbone settings. The following ablations examine the retrieval contributions of the architectural components, while the routing analyses examine concept-related expert organization.

Across repeated unseen-category splits, Table 3 shows that SG-MoE has the highest mean among task-specific configurations on NUS-WIDE and Pascal Sentence, with standard deviations comparable to those of the best-performing reference configurations.



(a) Concept-expert association measured by NMI and purity

(b) Aggregate concept-expert purity under semantic-input perturbations

(c) Routing entropy under semantic-input perturbations

Figure 2: Concept–expert organization and semantic-input sensitivity of SG-MoE routing on NUS-WIDE.

4.3 Ablation Studies

Table 4 reports component ablations on NUS-WIDE. Removing the Semantic Stream reduces I2T from 0.6067 to 0.5744 and T2I from 0.5987 to 0.5625, producing a 5.7% Avg mAP drop. Removing the Fine-Grained Stream has a smaller effect on I2T (0.5981) but lowers T2I to 0.5568, yielding a 4.2% Avg drop. Removing the Semantic Stream causes the larger I2T reduction, whereas removing the Fine-Grained Stream causes the larger T2I reduction, consistent with transferable semantic evidence from the former and local discrimination from the latter. Single-Expert Path retains the representative F- and S-stream expert paths activated during curriculum warm-up without activating the full sparse multi-expert architecture. Full SG-MoE raises Avg mAP from 0.5944 to 0.6027 relative to this existing architecture-level baseline. Because the two configurations are not matched for active parameter count or computation, this comparison does not isolate the effect of sparse routing or conditional computation alone.

Replacing modality-shared experts with modality-specific experts lowers I2T to 0.5924 and T2I to 0.5749, reducing Avg mAP by 3.2%. This result associates shared expert parameters with improved retrieval performance under the evaluated architecture, but does not by itself directly establish alignment of the internal semantic representations.

"w/o Semantic Router Learning" removes the training-time concept conditioning and auxiliary router supervision while retaining the MoE routers. Its I2T increases slightly and T2I decreases, leaving Avg mAP within 0.05% of Full SG-MoE. Thus, this ablation does not indicate a meaningful Avg mAP contribution from semantic router learning; the corresponding changes in concept–expert organization are examined in Section 4.4. Removing the curriculum decreases both retrieval directions and Avg mAP by about 1.0%.

4.4 Analysis of MLLM-Assisted Semantic Router Learning for Cross-Modal Representations

Figure 2 examines concept–expert organization and semantic-input sensitivity. **(a)** NMI and purity are reported for w/o Semantic Router Learning, Full SG-MoE under zero semantic input, and a random top-1 expert assignment baseline. Full SG-MoE improves modality-averaged NMI and purity over w/o Semantic Router Learning by 71.9% and 27.3%, respectively, indicating stronger concept–expert association at deployment. **(b)** With the model fixed, aggregate top-N purity is measured using the original encoder-generated concept distributions (Real), zero input (Inference zero), distributions reassigned across samples (Shuffled), and L1-normalized uniform random vectors (Random). Inference zero retains 93.9% of the Real-input purity, showing that the learned expert organization largely remains without external semantic input. **(c)** Routing entropy varies by at most 0.0003 across these settings, indicating that the purity differences cannot be explained by global changes in routing concentration. These results support concept-related expert organization

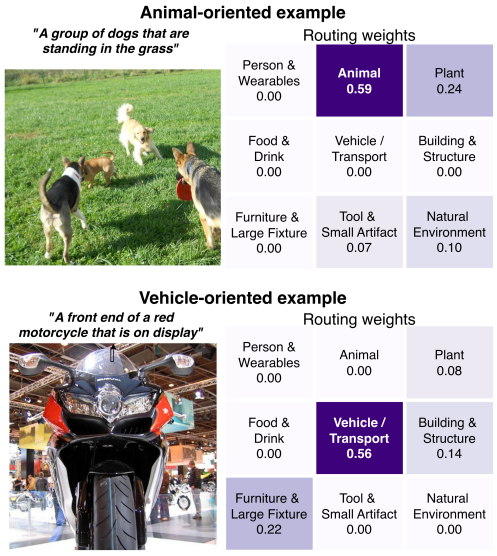


Figure 3: Concept-aligned routing examples. Darker cells indicate higher expert weights.

Table 4: Component ablation of cross-modal retrieval mAP on NUS-WIDE.

Variant	I2T	T2I	Avg
Full SG-MoE	0.6067	0.5987	0.6027
w/o Semantic Stream	0.5744	0.5625	0.5685
w/o Fine-Grained Stream	0.5981	0.5568	0.5775
Modality-Specific Experts	0.5924	0.5749	0.5837
w/o Semantic Router Learning	0.6106	0.5942	0.6024
Single-Expert Path	0.6035	0.5854	0.5944
w/o Curriculum	0.6026	0.5910	0.5969

Table 5: Computational cost and retrieval performance on NUS-WIDE.

Method	Params (M)	Train Time	Infer. Time	Avg mAP
DSCMR	0.52	∼5 min	∼0.65 s	0.4120
JDSH	0.61	∼5 min	∼0.68 s	0.4078
FGAN	1.15	∼30 min	∼0.82 s	0.4711
SG-MoE+TCR	2.84	∼20 min	∼1.03 s	0.5813
SG-MoE	2.84	∼20 min	∼1.12 s	0.6027

and routing interpretability rather than attributing the main retrieval gains to this module. All reported retrieval results follow the Inference zero protocol without external models.

Figure 3 provides sample-level routing examples. The dog and motorcycle scenes assign their highest weights to *Animal* (0.59) and *Vehicle / Transport* (0.56), respectively, while secondary experts capture visible context, consistent with the aggregate organization in Figure 2.

4.5 Computational Efficiency and Retrieval Performance

Table 5 summarizes computational cost and cross-modal retrieval performance on NUS-WIDE. Parameter counts cover only task-specific trainable components beyond the common VGG19-BN backbone. Train Time covers the default encoder-training and offline feature-export pipeline, excluding one-time ontology construction and CLIP-based concept-distribution generation; Infer. Time covers query matching or method-specific post-processing on available features, excluding encoder feature extraction and offline concept preparation. TCR adds no trainable parameters, so SG-MoE+TCR and SG-MoE share 2.84M task-specific parameters and an approximately 20-minute training/export pipeline, with measured matching or post-processing times of approximately 1.03s and 1.12s, respectively. SG-MoE without TCR improves Avg mAP from 0.5813 to 0.6027 and avoids test-time adaptation. The excluded offline stages preclude an end-to-end efficiency comparison.

5 Conclusion

We proposed SG-MoE: Shared Multi-Granularity MoE with MLLM-Assisted Semantic Router Learning for Cross-Modal Representation Learning and Zero-Shot Retrieval. We designed Dual-Stream Multi-Granularity MoE for Cross-Modal Representation Learning and Zero-Shot Retrieval to coordinate fine-grained and transferable semantic-level representations; established Cross-Modal Representation Alignment Based on Modality-Shared Experts to constrain internal cross-modal transformations; and developed MLLM-Assisted Semantic Router Learning for Cross-Modal Representations using an offline MLLM-assisted ontology, without external models at inference. Across four benchmarks, the dual-stream structure and modality-shared experts provide the main retrieval gains, while semantic router learning strengthens concept–expert associations.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and manuscript formatting. The authors reviewed all AI-assisted edits and take responsibility for the final content of this work.

REFERENCES

Mothilal Asokan, Kebin Wu, and Fatima Albreiki. FineLIP: Extending CLIP's reach via fine-grained alignment with longer text inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14495–14504, 2025.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32897–32912, 2022. doi: 10.52202/068431-2384.

Qizhou Chen, Chengyu Wang, Dakan Wang, Taolin Zhang, Wangyue Li, and Xiaofeng He. Lifelong knowledge editing for vision language models with low-rank mixture-of-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiqiang Luo, and Yantao Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval*, 2009.

Fartash Faghri, David J. Fleet, Jamie Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference*, 2018.

Rui Gao, Miaomiao Cheng, Xu Han, and Wei Song. High-order semantic alignment for unsupervised fine-grained image-text retrieval. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 8155–8165, 2024. URL https://aclanthology.org/2024.lrec-main.714/.

Shiping Ge, Zhiwei Jiang, Yafeng Yin, Cong Wang, Zifeng Cheng, and Qing Gu. Fine-grained alignment network for zero-shot cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(10):1–24, 2025. doi: 10.1145/3722223.

Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. CL-MoE: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Mark J. Huiskes and Michael S. Lew. The MIR Flickr retrieval evaluation. In *ACM International Conference on Multimedia Information Retrieval*, 2008.

Linglin Jing, Yuting Gao, Zhigang Wang, Wang Lan, Yiwen Tang, Weiyun Wang, Wenhai Wang, and Qingpei Guo. EvoMoE: Expert evolution in mixture of experts for multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22463–22471, 2026. doi: 10.1609/aaai.v40i27.39405.

Hyunji Lee, Luca Soldaini, Arman Cohan, Minjoon Seo, and Kyle Lo. RouterRetriever: Routing over a mixture of expert embedding models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11995–12003, 2025. doi: 10.1609/aaai.v39i11.33306.

Haobin Li, Peng Hu, Qianjun Zhang, Xi Peng, Xiting Liu, and Mouxing Yang. Test-time adaptation for cross-modal retrieval with query shift. In *International Conference on Learning Representations*, 2025. Spotlight.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-MoE: Scaling unified multimodal LLMs with mixture of experts. *arXiv preprint arXiv:2410.15017*, 2024.

Kaiyi Lin, Xing Xu, Lianli Gao, Zheng Wang, and Heng Tao Shen. Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

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*, pp. 740–755, 2014.

Bin Liu and Xiao-Ming Wu. Joint-modal distribution-based similarity hashing for cross-modal retrieval. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pp. 735–741, 2020.

Jiale Liu, Haoming Zhou, Yishu Liu, Bingzhi Chen, and Yuncheng Jiang. Cross modal fine-grained alignment via granularity-aware and region-uncertain modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23703–23711, 2026. doi: 10.1609/aaai.v40i28.39544.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with LIMoE: The language-image mixture of experts. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9564–9576, 2022. doi: 10.52202/068431-0695.

Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Man Luo, Sungduk Yu, Chendi Xue, and Vasudev Lal. Probing semantic routing in large mixture-of-expert models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18263–18278, 2025. doi: 10.18653/v1/2025.findings-emnlp.991. URL https://aclanthology.org/2025.findings-emnlp.991/.

Zhengxin Pan, Fangyu Wu, and Bailing Zhang. Fine-grained image-text matching by cross-modal hard aligning network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19275–19284, 2023.

Sumin Park and Noseong Park. How many experts are enough? towards optimal semantic specialization for mixture-of-experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24792–24800, 2026. doi: 10.1609/aaai.v40i29.39665.

Yang Qin, Yuan Sun, Xi Peng, Dezhong Peng, Joey Tianyi Zhou, Xiaomin Song, and Peng Hu. Robust semi-paired multimodal learning for cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24964–24972, 2026. doi: 10.1609/aaai.v40i30.39684.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using Amazon's Mechanical Turk. In Chris Callison-Burch and Mark Dredze (eds.), *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 139–147, Los Angeles, June 2010. Association for Computational Linguistics. URL https://aclanthology.org/W10-0721/.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Dianyi Wang, Siyuan Wang, Yitong Li, Duyu Tang, Xiaoyu Shen, Yikun Wang, Zejun Li, Xuanjing Huang, and Zhongyu Wei. MoIIE: Mixture of intra- and inter-modality experts for large vision language models. *arXiv preprint arXiv:2508.09779*, 2025.

Zheng Wang, Xing Xu, Jiwei Wei, Ning Xie, Yang Yang, and Heng Tao Shen. Semantics disentangling for cross-modal retrieval. *IEEE Transactions on Image Processing*, 33:2226–2237, 2024. doi: 10.1109/TIP.2024.3374111.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022.

Yang Yiyuan, Guodong Long, Michael Blumenstein, Xiubo Geng, Chongyang Tao, Tao Shen, and Daxin Jiang. Pre-training cross-modal retrieval by expansive lexicon-patch alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 12977–12987, 2024. URL https://aclanthology.org/2024.lrec-main.1136/.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Liangli Zhen, Peng Hu, Xu Wang, and Pengxi Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10394–10403, 2019.