MOVE: MULTIMODAL OPEN-WORLD VERIFICATION AND EXPANSION FOR GRAPH LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal graph learning faces a fundamental challenge: new classes may emerge after deployment, while models are trained with a fixed label space. Existing approaches typically detect unknown nodes and use LLMs to generate candidate class descriptions, but they do not determine whether existing classes are insufficient to cover these nodes or whether a generated class is reliable enough to expand the class space. Our empirical study reveals three challenges: multimodal information beyond individual modalities is required for unknown-node identification, LLM-generated class descriptions may not fully capture multimodal class characteristics, and directly adding candidate classes can introduce redundant categories. Based on these observations, we propose MOVE, a multimodal open-world class verification and expansion framework. MOVE identifies nodes that cannot be assigned to existing classes by jointly considering visual tokens, textual attributes, and graph context, leverages a multimodal LLM to generate candidate classes, and selectively expands the class space only when candidates are consistently supported by multimodal evidence without introducing unnecessary categories. Experiments demonstrate that MOVE achieves an average improvement of 2.73% across unknown recognition, open-domain annotation, and downstream graph learning tasks.

1 INTRODUCTION

Multimodal graphs provide a natural representation for real-world systems such as e-commerce platforms, social networks, and multimedia recommendation (Yan et al., 2025; Meng et al., 2025). However, these systems continuously evolve after deployment: new products, topics, and user interests may emerge beyond the classes observed during training. This motivates open-world multimodal graph learning (MM-OGL), where models need to identify nodes that cannot be assigned to existing classes and determine whether the class space should be expanded. Recent advances in open-world graph learning (OGL), such as OGA (Wen et al., 2025), move beyond unknown-node detection by assigning semantic labels to discovered nodes and incorporating newly identified classes into the model. Yet, for multimodal graphs, a natural-language class description alone is insufficient: the class of a node may also depend on visual tokens and its surrounding graph context, which may provide information unavailable from text alone. Therefore, the key problem is not simply generating a class name, but determining whether the visual, textual, and structural information of a node can still place it into an existing class, and whether a newly proposed class is sufficiently supported by these multimodal signals to justify expanding the class space.

To understand where this process fails, we conduct an empirical study following the OGL pipeline in Sec. 4. ① **Multimodal Assignment Failure.** The first challenge is determining whether a node truly belongs to an unseen class. A node may share similar visual or textual characteristics with known classes, while considering its visual tokens, textual tokens, and graph context together reveals that it cannot be assigned to any existing class (Sec. 4.1). This indicates that unknown discovery should consider more than similarity in a single modality. ② **Ambiguous Class Induction.** Once such nodes are identified, the next challenge is determining what class they represent. Different visual patterns, textual descriptions, and neighboring nodes may provide different clues, so a multimodal LLM can produce multiple plausible class descriptions (Sec. 4.2). Such a description is only a candidate: before being treated as a new class, it should be mapped back to the multimodal representation space and checked against the visual, textual, and structural information of the corresponding nodes.

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.

③ **Unreliable Class Expansion.** However, even a candidate supported by the current nodes should not be directly added to the model. It may only fit the observations used to generate it or unnecessarily split an existing class, causing redundant categories when the class space is expanded (Sec. 4.3).

These observations suggest that MM-OGL should be viewed as a continuous process from identifying nodes outside the current class space, to inferring what class they may belong to, and finally deciding whether that class should be added. The key is that each decision should use the same multimodal information rather than relying on a language label alone: visual and textual tokens together with graph context determine whether a node can be assigned to an existing class, while a proposed class must also be supported by these signals before it is incorporated.

Motivated by this perspective, we propose **MOVE** (**M**ultimodal **O**pen-world Class **V**erification and **E**xpansion), which couples the above three decisions in a unified framework. MOVE first performs **Uncertainty-Aware Multimodal Coding** to identify nodes that cannot be assigned to existing classes by jointly modeling visual tokens, textual tokens, and graph context. It then conducts **Evidence-Grounded Class Induction**, where a multimodal LLM proposes candidate class descriptions and each description is transformed into the shared multimodal space to determine whether it is supported by the corresponding nodes. Finally, **Selective Class Expansion** evaluates whether the supported candidate should be introduced into the class space, retaining only candidates that remain effective beyond the observations used to generate them while avoiding redundant classes and interference with existing ones.

Our Contributions. (1) ***New Perspective.*** We reveal a progressive failure chain in MM-OGL, from identifying nodes beyond existing classes, to ambiguous class annotation and unreliable class expansion. We formulate the problem as determining when the current class space becomes insufficient and when a new class should be introduced. (2) ***New Framework.*** We propose MOVE, a MM-OGL framework that unifies unknown-node identification, candidate class induction, and selective class expansion, with multimodal LLMs serving only as candidate generators rather than final decision makers. (3) ***SOTA Performance.*** Experiments on four multimodal attributed graph benchmarks and two downstream tasks show that MOVE achieves the best performance and improves unknown-concept semantic similarity by 7.62% on average.

2 Problem Formulation

We consider a multimodal attributed graph (MAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X}^v, \mathbf{X}^t)$, where each node is characterized by visual and textual attributes together with graph topology. During training, only a subset $\mathcal{V}_l \subset \mathcal{V}$ is labeled, whose labels belong to the known concept set $\mathcal{C}^K$. After deployment, the remaining nodes $\mathcal{V}_u = \mathcal{V} \setminus \mathcal{V}_l$ may contain both known and previously unseen concepts:

$$\mathcal{V}_u = \mathcal{V}_u^K \cup \mathcal{V}_u^U, \quad \mathcal{Y}(\mathcal{V}_u^K) \subseteq \mathcal{C}^K, \quad \mathcal{Y}(\mathcal{V}_u^U) \subseteq \mathcal{C}_t^U, \quad \mathcal{C}^K \cap \mathcal{C}_t^U = \emptyset. \tag{1}$$

Let $\mathcal{K}_t$ denote the concept knowledge available to the learner at stage t, initialized from the known concepts. Our goal goes beyond merely separating known and unseen nodes. Given $\mathcal{G}$ and $\mathcal{K}_t$, the learner should: (1) identify observations that cannot be explained by existing concepts while preserving known-class recognition; (2) infer coherent semantic concepts for the discovered unknown observations; and (3) selectively incorporate reliable concepts to obtain an updated knowledge state $\mathcal{K}_{t+1}$. Thus, open-world multimodal graph learning is formulated as a progressive process from *knowledge insufficiency* to *semantic discovery* and finally *knowledge evolution*.

3 Related Work

Open-World Graph Learning. Early open-world graph learning mainly focuses on rejecting nodes beyond the known label space. OpenWGL (Wu et al., 2020) models uncertainty for unseen-node recognition, followed by studies that improve OOD detection through predictive uncertainty (Stadler et al., 2021), neighborhood information, and structure-aware decision boundaries. Recent works further move beyond binary known–unknown separation: OpenIMA (Wang et al., 2024) and ORAL (Jin et al., 2024) explore novel class discovery, while OGA (Wen et al., 2025) extends the pipeline toward unknown recognition, semantic annotation, and model updating with LLM assistance. Despite this progress, existing methods still treat discovery, annotation, and updating as sepa-

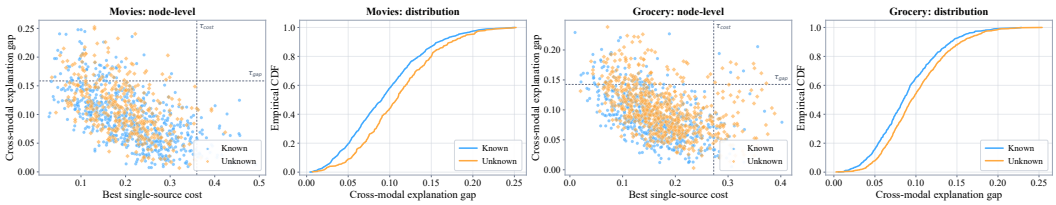


Figure 1: **Cross-modal explanation gaps.** The panels compare explanation gaps between known and unknown nodes with their cumulative distributions.

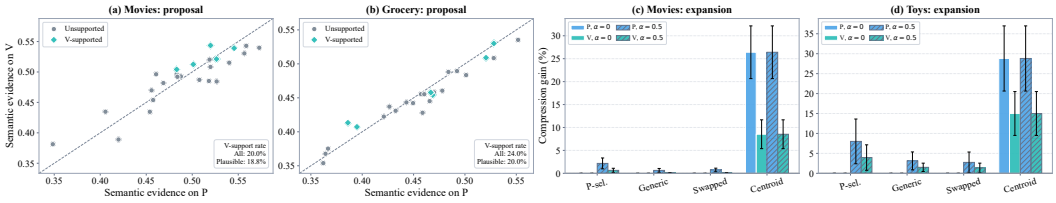


Figure 2: **Semantic proposal ambiguity and class expansion.** (a–b) Semantic evidence on proposal P and validation V sets; annotations give V-support rates. (c–d) Compression gains on P and V across adaptation settings. Error bars show standard errors across discovered regions.

rate stages with different criteria. MOVE instead unifies these decisions under a shared compression-based knowledge evolution principle.

Multimodal Graph Learning and Foundation Models. Multimodal graph learning has evolved from modality-specific encoding and graph fusion toward foundation-model-driven semantic reasoning. Recent benchmarks such as MAGB (Yan et al., 2025) systematically study multimodal graph understanding, while GraphGPT-O (Fang et al., 2025) and multimodal graph foundation models such as PLANET (Liu et al., 2026) improve cross-modal alignment and transferable representation learning. However, these methods mainly focus on improving the representation of existing semantic spaces and assume that concepts are predefined. MOVE further considers when the current semantic space becomes insufficient and verifies whether newly proposed concepts are worthy of being incorporated into knowledge.

4 Empirical Study

Before introducing MOVE, we empirically investigate the three observations behind our design: whether unknown concepts are jointly unexplained across modalities, whether semantic proposals are reliably grounded, and whether plausible concepts truly deserve knowledge expansion(see Appendix A).

4.1 Multimodal Assignment Failure

We first examine whether unknown concepts can be identified by individual modalities. For node i, we measure the gap between its best single-source match and the cost of jointly explaining visual, textual, and structural evidence:

$$G_i = \min_k \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} d(\mathbf{z}_i^r, \mathbf{c}_k^r) - \min_{r,k} d(\mathbf{z}_i^r, \mathbf{c}_k^r), \quad \mathcal{R} = \{v, t, g\}. \tag{2}$$

As shown in Fig. 1, **13.8%** of unknown nodes remain close to known concepts in at least one modality, but exhibit larger joint explanation gaps. This indicates that open-world novelty is better captured by multimodal explanation failure than isolated anomalies.

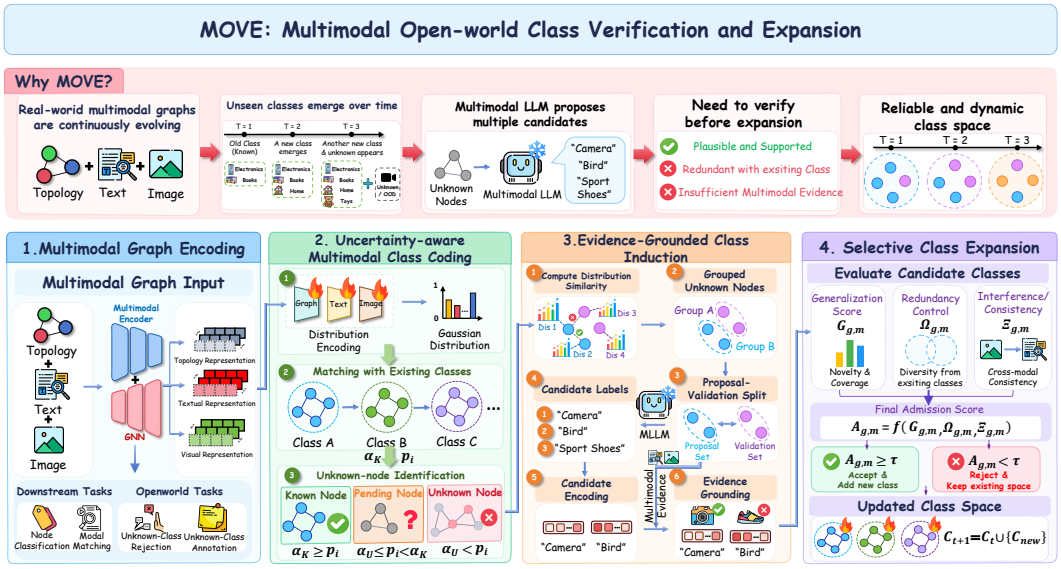


Figure 3: Overview of MOVE.

4.2 Ambiguous Class Induction

We next investigate whether semantic proposals can be directly obtained from unresolved observations. Given unknown regions, we ask an MLLM to generate candidate concepts and evaluate their support from multimodal evidence. Fig. 2(a–b) shows that although multiple concepts can be linguistically plausible, only **19.4%** receive consistent evidence support. This demonstrates that semantic generation alone is insufficient, motivating evidence-grounded hypothesis verification.

4.3 Unreliable Class Expansion

Finally, we examine whether improving local fitting is sufficient for knowledge expansion. For a candidate concept p, we compare its compression gain on the proposal set and independent evidence:

$$G^{P/V}(p) = R(S_g^{P/V}; \mathcal{K}) - R(S_g^{P/V}; \mathcal{K} \cup \{p\}). \tag{3}$$

As shown in Fig. 2(c–d), generic concepts can reduce local fitting error by **1.9%** due to increased capacity, but their gain decreases on independent evidence. Therefore, reliable knowledge evolution requires validating whether a concept provides transferable explanatory value rather than merely better fitting existing observations.

5 Methodology

5.1 Overview

As shown in Fig. 3, MOVE consists of three tightly coupled stages. **Uncertainty-Aware Multimodal Class Coding** evaluates whether existing classes can accommodate a node by jointly considering visual, textual, and structural information. **Evidence-Grounded Class Induction** uses a multimodal LLM to generate candidate classes for nodes that cannot be assigned to existing classes and verifies them with multimodal evidence. Finally, **Selective Class Expansion** determines whether a candidate class should be incorporated by evaluating its ability to characterize previously unresolved nodes while avoiding redundant categories and interference with existing classes. Details are applied in App.B.

5.2 Uncertainty-Aware Multimodal Class Coding

Motivation. Obs. 1 in Sec. 4.1 shows that a node may appear similar to known classes in an individual modality, while the joint visual, textual, and structural information prevents it from being

assigned to any existing class. Therefore, MOVE identifies unknown nodes by evaluating whether current classes can sufficiently represent their multimodal characteristics rather than relying on isolated modality similarity.

Distributional multimodal representation. To preserve uncertainty from heterogeneous modalities, MOVE represents each node as modality-specific distributions instead of deterministic embeddings:

$$q_i^r(\mathbf{z}) = \mathcal{N}\left(\boldsymbol{\mu}_i^r, \mathrm{diag}[(\boldsymbol{\sigma}_i^r)^2]\right), \quad (\boldsymbol{\mu}_i^r, \log \boldsymbol{\sigma}_i^r) = \phi_r(\mathbf{x}_i^r), \quad r \in \{v, t, g\}, \tag{4}$$

where $\mathbf{x}_i^v$, $\mathbf{x}_i^t$, and $\mathbf{x}_i^g$ denote visual tokens, textual attributes, and graph-context representations, respectively. Since different modalities may have different reliability, we assign adaptive weights:

$$\omega_i^r \propto \exp[-\mathcal{H}(q_i^r)/\tau_\omega], \quad \sum_r \omega_i^r = 1.$$

Given an existing class representation p_k, define

$$C_{ik} := \sum_r \omega_i^r D_{\mathrm{SKL}}(q_i^r \| p_k), \qquad D_i := \sum_{r<s} \omega_i^r \omega_i^s D_{\mathrm{B}}(q_i^r, q_i^s), \qquad E_{ik} := C_{ik} + \gamma D_i. \tag{5}$$

Here, C_{ik} measures class compatibility, whereas D_i measures cross-modal disagreement. MOVE uses them for class assignment and unknown rejection:

$$a_{ik} := \frac{\pi_k \exp(-C_{ik}/\tau_c)}{\sum_l \pi_l \exp(-C_{il}/\tau_c)}, \qquad \mathcal{R}_i(\mathcal{C}_t) := -\tau_c \log \sum_{k=1}^{K_t} \pi_k \exp(-C_{ik}/\tau_c) + \gamma D_i. \tag{6}$$

Because D_i is shared across classes, it cancels from a_{ik} and does not affect the preferred known class. It instead raises the rejection cost of nodes with inconsistent multimodal evidence.

The class representations and modality encoders are optimized on known nodes through

$$\mathcal{L}_{\mathrm{code}} := -\frac{1}{|\mathcal{V}_{\mathrm{tr}}^K|} \sum_{i \in \mathcal{V}_{\mathrm{tr}}^K} \log a_{iy_i}. \tag{7}$$

The distributional parameters receive gradients through C_{ik} and ω_i^r, whereas D_i and γ receive no direct signal from $\mathcal{L}_{\mathrm{code}}$. We fix γ before calibration and testing.

Unknown-node identification. MOVE compares each coding cost with a held-out known-class reference:

$$\hat{c}_i := \arg\max_k a_{ik}, \qquad p_i := \frac{1 + \sum_{s \in \mathcal{D}_K(\hat{c}_i)} \mathbb{I}[s \geq \mathcal{R}_i(\mathcal{C}_t)]}{1 + |\mathcal{D}_K(\hat{c}_i)|},$$
$$b_i := -\log_2 p_i, \qquad \mathrm{state}(i) := \begin{cases} \text{known}, & p_i > \alpha_K, \\ \text{pending}, & \alpha_U < p_i \leq \alpha_K, \\ \text{unknown}, & p_i \leq \alpha_U. \end{cases} \tag{8}$$

Here, $\mathcal{D}_K(\hat{c}_i)$ contains held-out known nodes assigned to predicted class $\hat{c}_i$. The cutoffs $\alpha_U < \alpha_K$ are fixed before testing. Only unknown nodes seed new-class discovery, while pending nodes provide supporting evidence.

5.3 Evidence-Grounded Class Induction

Motivation. Obs. 2 in Sec. 4.2 shows that nodes outside the current class space do not directly reveal their missing class. Although multimodal LLMs can generate plausible class descriptions, such descriptions may not fully reflect visual, textual, and structural characteristics. Therefore, MOVE treats LLM outputs as candidate class hypotheses and verifies them using multimodal evidence.

Multimodal evidence aggregation. For unknown nodes, MOVE first aggregates their multimodal information:

$$q_i(\mathbf{z}) = \frac{1}{Z_i} \prod_{r \in \{v,t,g\}} [q_i^r(\mathbf{z})]^{\omega_i^r}, \quad \{S_g\}_{g=1}^G = \mathrm{Cluster}\left(\{q_i \mid \mathrm{state}(i) = \mathrm{unknown}\}; D_{\mathrm{B}}\right). \tag{9}$$

Each group is divided into proposal and validation subsets:

$$S_g = S_g^P \cup S_g^V, \quad S_g^P \cap S_g^V = \emptyset.$$

The proposal subset is used for generating candidate classes, while the validation subset is reserved for evaluating whether a candidate generalizes beyond the observations used for generation.

Candidate class generation and grounding. For each proposal group, multimodal observations are organized as an evidence package $\mathcal{E}_g$. A multimodal LLM generates candidate class descriptions $\mathcal{H}_g = \{h_{g,m}\}_{m=1}^{M}$, which are encoded into the same semantic space as node representations as Gaussian class priors:

$$q_{g,m}^h = \mathcal{N}(\boldsymbol{\mu}_{g,m}^h, \boldsymbol{\Sigma}_{g,m}^h), \quad h_{g,m} \in \mathcal{H}_g. \tag{10}$$

To combine the semantic prior with observed evidence, MOVE first constructs a confidence-weighted, moment-matched group distribution $\bar{q}_g = \mathcal{N}(\bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g)$, which preserves both within-node uncertainty and between-node scatter (Appendix B). Let $\bar{\boldsymbol{\Lambda}}_g = \bar{\boldsymbol{\Sigma}}_g^{-1}$. We then perform scale-preserving precision pooling:

$$\boldsymbol{\Lambda}_{g,m}^* = \frac{\beta \boldsymbol{\Lambda}_{g,m}^h + \bar{\boldsymbol{\Lambda}}_g}{\beta + 1}, \boldsymbol{\mu}_{g,m}^* = \left(\beta \boldsymbol{\Lambda}_{g,m}^h + \bar{\boldsymbol{\Lambda}}_g\right)^{-1} \left(\beta \boldsymbol{\Lambda}_{g,m}^h \boldsymbol{\mu}_{g,m}^h + \bar{\boldsymbol{\Lambda}}_g \bar{\boldsymbol{\mu}}_g\right). \tag{11}$$

This normalization prevents the candidate covariance from shrinking merely as the proposal-group size increases. The grounded candidate class is evaluated according to its ability to improve class assignment:

$$u_{g,m} = \frac{1}{|S_g^P|} \sum_{i \in S_g^P} \left[\mathcal{R}_i(\mathcal{C}_t) - \mathcal{R}_i(\mathcal{C}_t \cup \{p_{g,m}^*\})\right], \quad \rho_{g,m} = \frac{\exp(u_{g,m}/\tau_h)}{\sum_j \exp(u_{g,j}/\tau_h)}. \tag{12}$$

The LLM provides candidate class semantics, while multimodal evidence determines whether the candidate is sufficiently supported.

5.4 Selective Class Expansion

Motivation. Obs. 3 in Sec. 4.3 shows that a candidate class that fits current unknown nodes may still introduce redundant categories. Therefore, MOVE evaluates candidates on unseen evidence before expanding the class space.

Cross-validated class improvement. For candidate class $p_{g,m}^*$, MOVE evaluates its generalization ability on the validation subset by measuring the reduction of class assignment cost:

$$G_{g,m}^{\mathrm{LCB}} = \frac{1}{|S_g^V|} \sum_{i \in S_g^V} \left[\mathcal{R}_i(\mathcal{C}_t) - \mathcal{R}_i(\mathcal{C}_t \cup \{p_{g,m}^*\})\right] - z_{1-\delta_{\mathrm{conf}}} \frac{\widehat{\sigma}_{g,m}}{\sqrt{|S_g^V|}}. \tag{13}$$

This validation process filters candidates that only fit the nodes used for generation.

Redundancy and interference control. To avoid unnecessary category expansion, MOVE considers class redundancy:

$$R_{g,m}^{\mathrm{red}} = \sum_{k=1}^{K_t} \pi_k \exp\left(-\frac{D_{\mathrm{B}}(p_{g,m}^*, p_k)}{\tau_\Omega}\right), C_{g,m}^{\mathrm{cmp}} = \frac{d_{\mathrm{eff}}}{2|S_g^P|} \log(1 + |S_g^P|), \Omega_{g,m} = R_{g,m}^{\mathrm{red}} + \lambda_{\mathrm{cmp}} C_{g,m}^{\mathrm{cmp}}. \tag{14}$$

We further measure whether a new class changes assignments of existing nodes:

$$\Xi_{g,m} = \frac{1}{|\mathcal{B}_{\mathrm{old}}|} \sum_{i \in \mathcal{B}_{\mathrm{old}}} D_{\mathrm{JS}}\left([\mathbf{a}_i^-, 0] \| \mathbf{a}_{i,g,m}^+\right). \tag{15}$$

The final admission score is:

$$\mathcal{A}_{g,m} = G_{g,m}^{\mathrm{LCB}} - \lambda_\Omega \Omega_{g,m} - \lambda_\Xi \Xi_{g,m}, \quad m^* = \arg\max_m \mathcal{A}_{g,m}. \tag{16}$$

The candidate class is incorporated only when it provides reliable improvement:

$$\mathcal{C}_{t+1} = \begin{cases} \mathcal{C}_t \cup \{(\pi_{g,m^*}, p_{g,m^*}^*)\}, & \mathcal{A}_{g,m^*} > 0, \\ \mathcal{C}_t, & \text{otherwise.} \end{cases} \tag{17}$$

Table 1: Known-class recognition in UCR.

Category	Method	Known-class Accuracy (↑)			
		Toys	Grocery	Reddit-S	Movies
OOD Detection	TopoOOD	85.39±0.54	84.43±0.54	97.83±0.10	59.74±0.61
	GNN-Safe	85.45±0.47	84.18±0.52	97.95±0.10	59.79±0.32
OOD Annotation	ORAL	82.55±0.80	83.51±0.31	96.35±0.14	56.63±0.74
	OpenWGL	78.17±0.64	71.46±1.55	97.34±0.70	52.34±0.82
MM-OOD	OGA	85.02±0.60	87.14±1.06	95.50±0.36	59.40±0.39
	AHM	82.51±0.73	81.14±0.90	96.67±0.62	55.74±0.71
Open World	CFC	63.89±2.85	51.62±6.56	79.07±4.08	38.88±4.60
	TopoNC	84.13±0.21	89.64±0.70	97.85±0.48	53.08±1.03
Ours	**MOVE**	**86.65±0.86**	**90.12±0.78**	**98.02±0.31**	**61.31±0.25**

Table 2: Unknown-node rejection in UCR.

Category	Method	Coverage / Precision (↑ / ↑)			
		Toys	Grocery	Reddit-S	Movies
OOD Detection	TopoOOD	10.68/54.28	8.61/53.92	12.30/54.26	9.75/22.14
	GNN-Safe	62.79/85.74	59.94/88.16	90.88/87.49	14.93/29.04
OOD Annotation	ORAL	30.67/74.23	35.99/82.15	41.80/77.10	15.05/29.52
	OpenWGL	72.90/88.25	62.93/88.83	95.49/91.36	22.39/39.54
MM-OOD	OGA	61.54/84.54	53.04/89.52	81.71/84.32	17.82/32.46
	AHM	36.33/76.86	30.26/79.64	89.74/83.65	16.26/31.09
Open World	CFC	60.70/80.58	47.98/84.09	89.57/88.79	26.01/28.93
	TopoNC	70.05/83.65	60.79/89.83	87.67/89.39	25.66/33.66
Ours	**MOVE**	**71.02/ 87.62**	**63.98/ 90.02**	**93.73/ 92.78**	**38.81/ 34.60**

Table 3: Semantic label quality in UCA, reported as G-to-U similarity (↑) / K-to-G similarity (↓).

Category	Method	Semantic Label Quality (G-to-U ↑ / K-to-G ↓)			
		Toys	Grocery	Reddit-S	Movies
OOD Annotation	ORAL	53.17±0.25 / 58.54±1.11	58.42±0.90 / 65.17±0.78	79.40±2.31 / 83.74±0.76	60.96±1.52 / 73.71±0.86
	OpenWGL	48.51±3.31 / 60.83±1.58	55.21±0.93 / 67.25±1.58	78.47±3.83 / 83.81±0.32	60.88±0.31 / 73.22±0.54
MM-OOD	OGA	53.86±0.62 / 58.22±0.86	58.78±0.92 / 63.73±0.66	79.46±1.86 / 85.97±1.08	60.01±0.38 / 72.80±0.58
	AHM	53.53±2.70 / 60.64±1.38	56.09±1.08 / 67.10±0.79	81.29±0.73 / 82.33±1.01	62.21±0.58 / 73.29±0.52
Open World	CFC	57.00±7.33 / 71.88±12.48	60.64±7.22 / 76.80±11.70	77.30±8.55 / 84.07±6.95	64.50±5.44 / 82.32±8.74
	TopoNC	56.29±6.74 / 73.13±12.00	59.88±5.87 / 77.75±12.86	77.31±8.61 / 83.08±6.98	65.22±5.16 / 83.01±10.08
Ours	**MOVE**	**64.39±3.11 / 53.84±5.18**	**65.78±2.70 / 53.76±2.17**	**86.09±0.78 / 81.86±0.48**	**78.36±2.32 / 72.47±3.13**

6 Experiments

In this section, we conduct a wide range of experiments and aim to answer the following questions: **Q1: Effectiveness.** Compared with other SOTA baselines, can MOVE achieve superior performance on both UCR and UCA? **Q2: Interpretability.** How do the core MDL components contribute to the overall success of MOVE? **Q3: Robustness.** How sensitive is MOVE to key hyperparameters, and how does it perform under severe cross-modal semantic inconsistency? **Q4: Efficiency.** What is the computational efficiency of MOVE, particularly regarding the MDL calculation and LLM-based semantic proposals? **Q5: Convergence and Scalability.** How do MOVE and competing methods behave as training progresses, and how does MOVE scale with increasing graph and category sizes?

6.1 Experimental Setup

We evaluate MOVE on four multimodal attributed graphs, including Toys, Grocery, Movies, and Reddit-S (Yan et al., 2025). We compare MOVE with representative open-world graph learning baselines, including OOD detection methods, OOD annotation methods, multimodal OOD methods and open-world methods. The evaluation covers the complete open-world pipeline, including unknown-class rejection (UCR), unknown-class annotation (UCA), downstream graph-level and modality-level task. Detailed descriptions are provided in Appendix D.

6.2 Performance Comparison (Q1)

UCR Performance. Tables 1 and 2 report known-class recognition and unknown-node rejection. In all tables, red, blue, and orange indicate the best, second-best, and third-best results, respectively. MOVE achieves the highest known-class Accuracy on all datasets, exceeding the strongest baseline by up to 1.52 percentage points, and ranks among the top two in UCR Coverage and Precision throughout. This indicates a favorable balance between known- and unknown-class performance.

UCA Performance. For semantic label quality, "K", "U", and "G" denote known, held-out unknown, and generated labels, respectively. As shown in Table 3, MOVE consistently achieves the highest G-to-U similarity and the lowest K-to-G similarity across all four datasets. This indicates that its generated concepts accurately capture unknown-class semantics while remaining distinguishable from the known label space.

Table 4: Downstream performance after UCA (%, ↑) on NC and MM (%, ↑).

Category	Method	Toys		Grocery		Reddit-S		Movies	
		NC (Acc.)	MM	NC (Acc.)	MM	NC (Acc.)	MM	NC (Acc.)	MM
Reference	Lower bound	39.44±0.55	-	32.38±0.67	-	39.49±1.11	-	39.57±0.67	-
OOD Annotation	ORAL	41.68±1.98	79.52±2.69	32.05±0.79	76.74±2.89	53.57±2.94	68.16±6.32	41.70±0.68	49.25±9.74
	OpenWGL	39.70±0.72	11.32±2.69	29.85±0.26	22.99±6.84	43.09±2.87	55.09±1.44	42.26±0.51	17.93±2.17
MM-OOD	OGA	42.57±0.55	72.70±3.42	36.74±1.21	83.76±4.38	47.38±4.33	82.61±1.35	41.81±0.60	38.77±9.90
	AHM	38.58±0.75	66.88±6.41	29.57±0.67	32.26±1.77	50.24±1.35	55.15±8.36	42.16±0.55	29.82±4.05
Open World	CFC	36.18±2.15	74.12±1.23	36.36±3.20	81.02±3.44	48.04±0.84	81.86±0.71	45.46±0.82	44.58±0.97
	TopoNC	31.24±2.25	73.67±0.67	33.72±1.02	82.97±0.60	48.48±0.74	81.83±1.03	45.81±1.04	46.27±3.66
Ours	**MOVE**	**46.65±1.74**	**76.69±1.64**	**37.60±1.40**	**84.87±3.25**	**57.27±3.02**	**86.54±3.93**	**46.80±0.72**	**50.95±2.46**
Reference	Upper bound	72.82±0.36	-	82.13±0.66	-	94.60±0.41	-	53.90±0.82	-

Table 5: Ablation study of MOVE. UCR reports Coverage/Precision, NC reports post-annotation GCN node-classification Accuracy (%).

Variants	Toys		Grocery		Movies		Reddit-S	
	UCR (Cov./Prec.)	NC (Acc.)	UCR (Cov./Prec.)	NC (Acc.)	UCR (Cov./Prec.)	NC (Acc.)	UCR (Cov./Prec.)	NC (Acc.)
w/o CDM	70.65 / 80.68	44.66	62.34 / 88.73	32.49	36.89 / 29.79	45.59	92.31 / 91.64	52.45
w/o HCV	69.65 / 80.23	40.01	62.39 / 88.85	32.32	35.76 / 30.33	44.29	91.74 / 92.22	38.99
w/o EG	71.90 / 80.56	49.55	60.26 / 87.65	32.49	37.96 / 29.35	45.97	91.68 / 91.97	38.99
w/o GCE	68.39 / 80.55	44.82	62.25 / 88.78	36.77	34.48 / 30.42	45.86	86.23 / 92.42	53.12
w/o CSU	-/-	40.01	-/-	32.32	-/-	44.84	-/-	39.03
MOVE (full)	**71.02 / 87.62**	**46.65**	**63.98 / 90.02**	**37.60**	**38.81 / 34.60**	**46.80**	**93.73 / 92.78**	**57.27**

Performance on Downstream Tasks. We evaluate downstream utility through node classification (NC) and zero-shot modality matching (MM). For NC, the lower and upper bounds use no unknown-label annotations and ground-truth unknown labels, respectively. As shown in Table 4, MOVE achieves the best node classification performance on all four datasets. For modality matching, it ranks first on three datasets and second on Toys. These results indicate that the generated concepts support both node classification and cross-modal alignment.

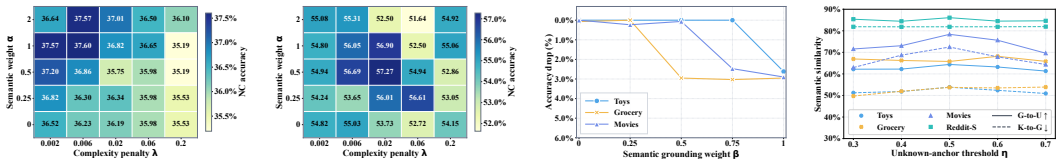
6.3 Ablation Analysis (Q2)

We evaluate five functional components of MOVE: Cross-Modal Disagreement Modeling (CDM), Held-out Candidate Validation (HCV), Evidence Grounding (EG), Graph-Context Encoding (GCE), and Class-Space Update (CSU). As shown in Table 5, CDM and GCE mainly support UCR, whereas HCV and CSU are particularly important for downstream learning. Removing EG decreases average UCR precision and post-annotation GCN accuracy by 3.87 and 5.33 percentage points, respectively. Detailed ablation settings are provided in Appendix D.

6.4 Robustness Analysis (Q3)

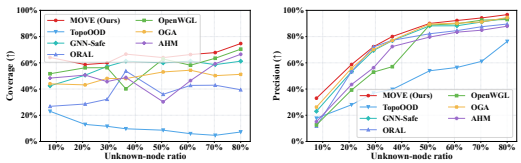
Hyperparameter Sensitivity. Fig. 4(a)–(d) show that MOVE remains stable across different settings of complexity penalties (λ, α), semantic grounding weight β, and unknown-anchor threshold η. Moderate regularization achieves a balance between accepting useful new classes and avoiding redundant expansion. The effect of β is dataset dependent, while the default $\eta = 0.5$ provides a robust trade-off between G-to-U similarity and K-to-G similarity.

Robustness to Increasing Openness. We increase the unknown-node ratio from 10% to 80% on Grocery and compare MOVE with representative open-world baselines under identical settings. As shown in Fig. 5, MOVE consistently achieves strong Coverage and Precision across different openness levels, with stable unknown-node rejection performance even when known-class supervision becomes increasingly limited. These results demonstrate the robustness of MOVE under challenging open-world scenarios.



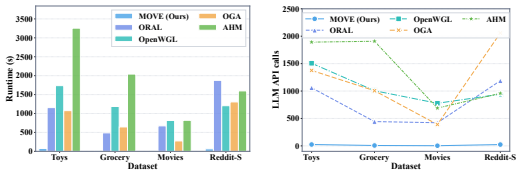
(a) Grocery: λ–α. (b) Reddit-S: λ–α. (c) Sensitivity to β. (d) Sensitivity to η.

Figure 4: **Hyperparameter robustness.** (a)–(b) NC accuracy versus λ and α. (c) Relative accuracy drop versus β. (d) Semantic label quality versus threshold η.



(a) Coverage. (b) Precision.

Figure 5: **Robustness to graph openness.**



(a) Runtime. (b) LLM API calls.

Figure 6: **Efficiency and computational cost.**

6.5 Efficiency Comparison (Q4)

To answer Q4, Fig. 6 compares the annotation-stage efficiency of MOVE with annotation-capable baselines in terms of wall-clock time and LLM API calls. MOVE achieves the lowest runtime across all datasets, being $14.1\times$–$18.1\times$ faster than the fastest baseline while reducing LLM calls from hundreds or thousands to only 3–25. These improvements result from the MDL/UCR gate and compact candidate grouping, which avoid unnecessary LLM queries by validating only coherent candidate regions before concept expansion.

6.6 Convergence and Scalability Analysis (Q5)

To answer Q5, we evaluate convergence over training and scalability with graph and category sizes. MOVE achieves the highest post-annotation NC accuracy at every checkpoint on Toys and Reddit-S, whereas the baselines fluctuate more or converge to lower levels. As the retained-node ratio increases, runtime and peak GPU memory grow moderately but remain below 4.2 seconds and 1.5 GiB at full scale. Although accuracy generally decreases with more classes, MOVE retains 37.60%–57.27% NC accuracy under the complete class configurations. Detailed results are provided in Appendix D.

7 Conclusion

In this paper, we study open-world multimodal graph learning from the perspective of progressive knowledge evolution. Through empirical analysis, we reveal that unknown concepts require joint multimodal explanation, LLM-generated concepts need evidence-based verification, and naive expansion may introduce redundant categories. Based on these observations, we propose MOVE, a unified framework that identifies unresolved observations through uncertainty-aware multimodal coding, induces candidate concepts with multimodal LLMs and evidence grounding, and selectively expands the class space with reliable and non-redundant concepts. Extensive experiments demonstrate that MOVE effectively improves unknown-class recognition, semantic annotation, and downstream graph learning while maintaining a reliable and dynamic evolving class space.

AI Usage Disclosure

Generative AI tools were used during the preparation of this manuscript for limited purposes, including language refinement, literature retrieval and discovery, and research ideation support. Specifically, AI tools assisted with improving writing clarity, organizing related work, exploring potential

research directions, and discussing possible methodological designs and experimental considerations.

Reproducibility Statement

We provide an anonymized implementation and experiment configurations at `https://anonymous.4open.science/r/MOVE-305E/`. The main paper describes the complete MOVE pipeline, learning objectives, and evaluation protocol. The appendix provides detailed empirical diagnostic protocols, multimodal feature construction, candidate-generation and verification procedures, training and deployment details, theoretical analyses, dataset splits, baseline adaptations, evaluation metrics, ablation settings, and additional robustness, efficiency, convergence, and scalability results. The released repository excludes the raw MAGB datasets, third-party pretrained checkpoints, and credentials for the external Qwen3-VL-Flash service, but documents the required data format and external model dependencies.

References

Tianyi Bao, Qitian Wu, Zetian Jiang, Yiting Chen, Jiawei Sun, and Junchi Yan. Graph out-of-distribution detection goes neighborhood shaping. In *ICML*, pp. 2923–2943, 2024.

Christos Constantinou, Georgios Ioannides, Aman Chadha, Aaron Elkins, and Edwin Simpson. Out of distribution detection with attention head masking for multimodal document classification. *Scientific Reports*, 16(1):2449, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yi Fang, Bowen Jin, Jiacheng Shen, Sirui Ding, Qiaoyu Tan, and Jiawei Han. Graphgpt-o: Synergistic multimodal comprehension and generation on graphs. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19467–19476. IEEE, 2025.

Yucheng Jin, Yun Xiong, Juncheng Fang, Xixi Wu, Dongxiao He, Xing Jia, Bingchen Zhao, and Philip S Yu. Beyond the known: Novel class discovery for open-world graph learning. In *International Conference on Database Systems for Advanced Applications*, pp. 117–133. Springer, 2024.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Sicheng Liu, Xunkai Li, Daohan Su, Ru Zhang, Hongchao Qin, Ronghua Li, and Guoren Wang. Toward effective multimodal graph foundation model: A divide-and-conquer based approach. *arXiv preprint arXiv:2602.04116*, 2026.

Xueqi Ma, Xingjun Ma, Sarah Monazam Erfani, Danilo Mandic, and James Bailey. Coarse-to-fine open-set graph node classification with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36714–36722, 2026.

Fanshen Meng, Zhenhua Meng, Ru Jin, Rongheng Lin, and Budan Wu. Doge: Llms-enhanced hyper-knowledge graph recommender for multimodal recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 12399–12407, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, and Stephan Günnemann. Graph posterior network: Bayesian predictive uncertainty for node classification. *Advances in Neural Information Processing Systems*, 34:18033–18048, 2021.

Yanling Wang, Jing Zhang, Lingxi Zhang, Lixin Liu, Yuxiao Dong, Cuiping Li, Hong Chen, and Hongzhi Yin. Open-world semi-supervised learning for node classification. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 2723–2736. IEEE, 2024.

Yanzhe Wen, Xunkai Li, Qi Zhang, Zhu Lei, Guang Zeng, Rong-Hua Li, and Guoren Wang. When llms meet open-world graph learning: a new perspective for unlabeled data uncertainty. *arXiv preprint arXiv:2505.13989*, 2025.

Man Wu, Shirui Pan, and Xingquan Zhu. Openwgl: Open-world graph learning. In *2020 IEEE international conference on data mining (icdm)*, pp. 681–690. IEEE, 2020.

Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. Energy-based out-of-distribution detection for graph neural networks. *arXiv preprint arXiv:2302.02914*, 2023.

Xuanzhi Xi, Zhong Zhang, Hongliang Wang, Tongze Zhang, hongchao hu, Qinli Yang, and Junming Shao. Topology-aware neural collapse for generalized category discovery on graphs. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5579–5590, 2026.

Hao Yan, Chaozhuo Li, Jun Yin, Zhigang Yu, Weihao Han, Mingzheng Li, Zhengxin Zeng, Hao Sun, and Senzhang Wang. When graph meets multimodal: benchmarking and meditating on multimodal attributed graph learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5842–5853, 2025.

A Empirical Study in Detail

This section specifies the diagnostic study in Sec. 4. The study is designed to isolate the three failure modes that motivate MOVE; it is not used to train MOVE or select its hyperparameters. Unless stated otherwise, all thresholds are calibrated using known-class data and fixed before candidate evaluation on independent unknown-class evidence.

A.1 Shared Representations and Diagnostic Protocol

Multimodal representations. We use the released CLIP (Radford et al., 2021) ViT-L/14 (Dosovitskiy et al., 2020) visual and textual features and normalize each feature to unit length, yielding $\mathbf{z}_i^v$ and $\mathbf{z}_i^t$. To obtain a topology-conditioned representation without introducing an additional learned model, we first form $\widetilde{\mathbf{z}}_i = \mathrm{norm}(\mathbf{z}_i^v + \mathbf{z}_i^t)$ and aggregate one-hop incoming neighbors with a self-loop:

$$\mathbf{z}_i^g = \mathrm{norm}\left(\frac{\widetilde{\mathbf{z}}_i + \sum_{j:(j,i)\in\mathcal{E}} \widetilde{\mathbf{z}}_j}{1 + |\{j : (j,i) \in \mathcal{E}\}|}\right). \tag{18}$$

The evidence sources are therefore $\mathcal{R} = \{v, t, g\}$, corresponding to image, text, and graph context. Cosine distance is denoted by $d(\mathbf{a}, \mathbf{b}) = 1 - \mathbf{a}^\top \mathbf{b}$ for unit vectors. The modality-specific prototype of each known class is estimated using its training nodes:

$$\mathbf{c}_k^r = \mathrm{norm}\left(\frac{1}{|\mathcal{T}_k|} \sum_{i\in\mathcal{T}_k} \mathbf{z}_i^r\right), \qquad r \in \mathcal{R}. \tag{19}$$

Proposal and validation evidence. For Obs. 2 and Obs. 3, held-out unknown nodes in the validation and test splits form disjoint proposal and validation pools, denoted by P and V, respectively. Unknown-class identities are not used for region discovery, candidate generation, candidate selection, threshold calibration, or support decisions; they are accessed only for post-hoc analysis.

We fuse the three evidence sources as

$$\mathbf{f}_i = \mathrm{norm}(\mathbf{z}_i^v + \mathbf{z}_i^t + \mathbf{z}_i^g) \tag{20}$$

and cluster the proposal pool into five unresolved regions. Central proposal nodes from each region form the candidate-generation subset P_g, while nodes in V are assigned to the corresponding frozen regions to form the independent validation subset V_g. This construction separates the evidence used to generate a candidate from the evidence used to test its generalization.

A.2 OBS. 1: CROSS-MODAL EXPLANATION GAP

For node i, the best joint explanation by the known dictionary and the best explanation available from any individual evidence source are, respectively,

$$C_i^{\text{joint}} = \min_k \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} d(\mathbf{z}_i^r, \mathbf{c}_k^r), \qquad C_i^{\text{single}} = \min_{r \in \mathcal{R}, k} d(\mathbf{z}_i^r, \mathbf{c}_k^r). \tag{21}$$

The cross-modal explanation gap reported in Fig. 1 is

$$G_i = C_i^{\text{joint}} - C_i^{\text{single}}. \tag{22}$$

A node is operationally defined as *single-source familiar but jointly unexplained* when

$$C_i^{\text{single}} \leq \tau_{\text{fam}}, \qquad G_i > \tau_{\text{gap}}, \tag{23}$$

where τ_{fam} and τ_{gap} are calibrated using known-class nodes. This avoids selecting an absolute distance scale using unknown labels.

Across all four datasets, this event contains 11.7% of unknown test nodes and 9.9% of known test nodes. The corresponding unknown/known rates are 13.7%/10.0% on Toys, 13.4%/9.4% on Grocery, 15.2%/10.1% on Movies, and 9.9%/6.2% on Reddit-S. The higher unknown-node rate is consistently observed across the four datasets, showing that an observation may appear familiar under an individual modality while remaining incompatible with the known class space when all evidence sources are considered jointly. The main text visualizes Movies and Grocery for clarity.

A.3 OBS. 2: MULTIMODAL EVIDENCE AND LINGUISTIC PLAUSIBILITY

Candidate generation. For each unresolved region, Qwen3-VL-Flash receives representative textual and visual evidence together with the names of known classes and generates five concise candidate descriptions. Neither V_g nor the identities of unknown classes are provided to the MLLM. Candidate descriptions are encoded using the CLIP ViT-L/14 text encoder and unit-normalized, yielding the semantic vector $\mathbf{q}_{g,m}$ for candidate m. Graph context is withheld from candidate generation and is used only as an additional source of evidence during verification.

Linguistic plausibility. Linguistic plausibility is assessed by the MLLM using proposal evidence only and is recorded before multimodal verification. A candidate is considered highly plausible when it provides a coherent description of the displayed proposal evidence. This judgment measures linguistic plausibility rather than correctness or support from independent evidence.

Operational definition of potential semantic hallucination. In this work, we use *potential semantic hallucination* to denote a candidate class description that appears linguistically plausible given the proposal evidence but is not consistently supported by independent multimodal evidence. Specifically, a highly plausible candidate is regarded as potentially hallucinated if it fails at least one of the support conditions in Eq. 27. This is an evidence-based operational definition rather than a claim of factual incorrectness: a failed candidate may be semantically inappropriate, overly broad, specific to the proposal observations, or insufficiently supported by the available evidence. Unknown-class identities are used only for post-hoc evaluation and are not involved in this determination.

Multimodal evidence support. For split $S \in \{P, V\}$, we summarize each evidence source in region g and measure its alignment with candidate m:

$$\overline{\mathbf{z}}_{g,S}^r = \text{norm}\left(\frac{1}{|S_g|} \sum_{i \in S_g} \mathbf{z}_i^r\right), \qquad a_{g,m,S}^r = (\overline{\mathbf{z}}_{g,S}^r)^\top \mathbf{q}_{g,m}. \tag{24}$$

The average multimodal alignment is

$$a_{g,m,S} = \frac{1}{3} \sum_{r \in \mathcal{R}} a_{g,m,S}^r. \tag{25}$$

The source-specific support threshold is calibrated exclusively using known validation nodes. Let $\mathcal{A}_r^+$ contain the alignments between each known-class name and the matching class evidence under source r, and let $\mathcal{A}_r^-$ contain the corresponding mismatched alignments. We define

$$\tau_r = \frac{1}{2}\left[Q_{0.25}(\mathcal{A}_r^+) + Q_{0.75}(\mathcal{A}_r^-)\right]. \tag{26}$$

A candidate is consistently supported on independent evidence if

$$\begin{aligned} a^r_{g,m,V} &\geq \tau_r && \text{for every } r \in \mathcal{R}, \\ \mathrm{rank}_{m'}(a^r_{g,m',V}) &\leq 2 && \text{for every } r \in \mathcal{R}, \\ a_{g,m,V} &\geq a_{g,m,P} - \epsilon, && \epsilon = 0.015. \end{aligned} \tag{27}$$

These conditions require the candidate to receive sufficient support from all three evidence sources, remain among the strongest alternatives under each source, and avoid a substantial proposal-to-validation degradation. The rules are fixed before candidate evaluation and do not use unknown-class identities.

In the Movies and Grocery results, 6 of the 31 highly plausible candidates satisfy all three conditions, yielding the reported support rate of 19.4%. The corresponding support rates within the highly plausible subset are 18.75% on Movies and 20.00% on Grocery. These results show that linguistic plausibility does not necessarily imply consistent support from multimodal evidence or generalization beyond the observations used for candidate generation. Movies and Grocery are visualized in the main text, while the same evaluation protocol is applied to all four datasets.

A.4 Obs. 3: Proposal Fit and Independent Expansion Evidence

Obs. 3 examines whether reducing reconstruction cost on proposal evidence is sufficient to justify class-space expansion. For each unresolved region, we compare four candidate constructions:

1. P**-selected**: the MLLM candidate with the largest mean proposal alignment $a_{g,m,P}$;
2. **Generic**: a deliberately broad dataset-level description;
3. **Swapped**: a candidate transferred from another unresolved region of the same dataset; and
4. **Centroid**: a candidate constructed directly from the modality-specific proposal centroids without natural-language semantics.

These controls respectively represent proposal-based selection, overly broad semantics, region-mismatched semantics, and direct local fitting.

For each language-based construction, the text vector is initially shared across evidence sources, $\mathbf{h}^r_{g,0} = \mathbf{q}_g$. We additionally evaluate a locally adapted version:

$$\mathbf{h}^r_{g,\alpha} = \mathrm{norm}\left[(1-\alpha)\mathbf{q}_g + \alpha \overline{\mathbf{z}}^r_{g,P}\right], \qquad \alpha \in \{0, 0.5\}. \tag{28}$$

The Centroid construction is already fully determined by proposal evidence and is therefore unchanged under this comparison.

Let the known-dictionary reconstruction cost be

$$C_i(\mathcal{K}) = \min_{k \in \mathcal{K}} \sum_{r \in \mathcal{R}} d(\mathbf{z}^r_i, \mathbf{c}^r_k). \tag{29}$$

After tentatively adding candidate $h = (\mathbf{h}^v, \mathbf{h}^t, \mathbf{h}^g)$, the reconstruction cost becomes

$$C_i(\mathcal{K} \cup \{h\}) = \min\left\{C_i(\mathcal{K}), \sum_{r \in \mathcal{R}} d(\mathbf{z}^r_i, \mathbf{h}^r)\right\}. \tag{30}$$

For $S \in \{P, V\}$, the relative compression gain is

$$\Delta^S_g(h) = 100 \times \frac{\frac{1}{|S_g|}\sum_{i \in S_g} C_i(\mathcal{K}) - \frac{1}{|S_g|}\sum_{i \in S_g} C_i(\mathcal{K} \cup \{h\})}{\frac{1}{|S_g|}\sum_{i \in S_g} C_i(\mathcal{K})}. \tag{31}$$

The same candidate is evaluated on both the proposal region P_g and the corresponding independent validation region V_g. In the Movies and Toys results, the locally adapted Generic control obtains an average compression gain of 1.9% on proposal evidence but only 0.8% on independent validation evidence. The average proposal-to-validation gain reductions are 6.36 percentage points on Movies and 5.61 percentage points on Toys. Error bars report standard errors across the five unresolved regions within each dataset.

These results show that additional representational capacity may improve proposal-set reconstruction even when the candidate does not provide transferable explanatory value. Therefore, proposal-set improvement alone cannot serve as a reliable class-admission rule, motivating the independent validation and selective-expansion mechanism in MOVE. Movies and Toys are visualized in the main text, while the same comparison is conducted on all four datasets.

B Methodology in Details

This section provides the implementation details of MOVE and clarifies the complete data flow shown in Fig. 3. The framework contains a multimodal graph encoding front-end followed by three core stages: **Uncertainty-Aware Multimodal Class Coding**, **Evidence-Grounded Class Induction**, and **Selective Class Expansion**. The encoding front-end converts raw image, text, and graph topology into modality-specific node representations. The three MOVE stages then determine whether a node can be assigned to the current class space, infer candidate classes for unresolved nodes, and selectively incorporate only those candidates that remain reliable under independent evidence.

B.1 Multimodal Graph Encoding

Input representation. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X}^v, \mathbf{X}^t)$ denote a multimodal attributed graph with adjacency structure $\mathbf{A}$. Each node i is associated with visual input $\mathbf{x}_i^v$ and textual input $\mathbf{x}_i^t$. We first encode the two node-level modalities independently:

$$\mathbf{e}_i^v = f_v(\mathbf{x}_i^v), \quad \mathbf{e}_i^t = f_t(\mathbf{x}_i^t), \tag{32}$$

where f_v and f_t denote the visual and textual encoders, respectively. Both representations are projected to a shared d-dimensional latent space before subsequent processing.

Topology-conditioned graph context. The graph branch is not treated as an independent raw modality. Instead, visual and textual features are first fused into an initial node representation,

$$\mathbf{h}_i^{(0)} = f_{\text{fuse}}\left([\mathbf{e}_i^v \| \mathbf{e}_i^t]\right), \tag{33}$$

which is propagated over the observed graph topology:

$$\mathbf{e}_i^g = \text{GNN}\left(\mathbf{A}, \{\mathbf{h}_j^{(0)}\}_{j\in\mathcal{V}}\right)_i. \tag{34}$$

Here, $\mathbf{e}_i^g$ represents topology-conditioned multimodal context rather than topology alone. MOVE therefore retains three complementary representations for each node: $\mathbf{e}_i^v$ describing its visual content, $\mathbf{e}_i^t$ describing its textual attributes, and $\mathbf{e}_i^g$ describing its neighborhood-aware graph context. These representations form the input to uncertainty-aware class coding.

B.2 Uncertainty-Aware Multimodal Class Coding

Modality-specific distributional encoding. A deterministic embedding does not indicate how reliable a modality is for a particular node. MOVE therefore maps the three representations to separate distributions. Each modality $r \in \{v, t, g\}$ is equipped with an independent trainable distribution head ϕ_r:

$$\boldsymbol{\mu}_i^r, \mathbf{s}_i^r := \phi_r(\mathbf{e}_i^r), \quad \boldsymbol{\sigma}_i^r := \sigma_{\min} + \text{softplus}(\mathbf{s}_i^r), \quad q_i^r(\mathbf{z}) := \mathcal{N}\left(\boldsymbol{\mu}_i^r, \text{diag}[(\boldsymbol{\sigma}_i^r)^2]\right). \tag{35}$$

The variance floor $\sigma_{\min} > 0$ prevents numerical collapse. The mean $\boldsymbol{\mu}_i^r$ represents the semantic location of modality r, while $\boldsymbol{\sigma}_i^r$ is a learned dispersion parameter that controls its relative reliability. We do not interpret $\boldsymbol{\sigma}_i^r$ as an independently calibrated predictive variance. It is learned through the class-compatibility objective and the reliability weights defined below.

Uncertainty-aware reliability. For a diagonal Gaussian, we compute its differential entropy as

$$\mathcal{H}(q_i^r) := \frac{1}{2}\sum_{d=1}^{D}\log\left(2\pi e(\sigma_{i,d}^r)^2\right), \quad \omega_i^r := \frac{\exp[-\mathcal{H}(q_i^r)/\tau_\omega]}{\sum_{s\in\{v,t,g\}}\exp[-\mathcal{H}(q_i^s)/\tau_\omega]}. \tag{36}$$

A modality with lower learned dispersion receives a larger reliability weight. The node-specific weights satisfy $\sum_r \omega_i^r = 1$, allowing MOVE to reduce the contribution of noisy or ambiguous modalities without manually selecting a dominant information source.

Existing-class representations. Each existing class k is represented in the same latent space by a diagonal Gaussian

$$p_k(\mathbf{z}) := \mathcal{N}\left(\boldsymbol{\mu}_k, \mathrm{diag}[\boldsymbol{\sigma}_k^2]\right), \quad k = 1, \ldots, K_t. \tag{37}$$

The class distributions are initialized from known-class training observations and optimized jointly with the encoding and distributional heads through the class-dependent coding objective in Eq. 7. The class prior π_k reflects the relative support of class k in the known training data and satisfies $\sum_k \pi_k = 1$.

Multimodal class matching. For each node–class pair (i, k), MOVE separates class compatibility from node-wise cross-modal disagreement:

$$\begin{aligned} C_{ik} &:= \underbrace{\sum_{r\in\{v,t,g\}} \omega_i^r D_{\mathrm{SKL}}(q_i^r \| p_k)}_{\text{class compatibility}}, \\ D_i &:= \underbrace{\sum_{r<s} \omega_i^r \omega_i^s D_{\mathrm{B}}(q_i^r, q_i^s)}_{\text{cross-modal disagreement}}, \qquad E_{ik} := C_{ik} + \gamma D_i. \end{aligned} \tag{38}$$

The compatibility C_{ik} depends on class k, whereas D_i is shared across all classes for the same node. Consequently, disagreement cannot determine which known class is preferred. It instead supplies a class-independent penalty when the modalities of a node provide inconsistent evidence.

Class assignment and coding cost. The class-assignment probabilities depend only on the class-specific compatibility:

$$a_{ik} := \frac{\pi_k \exp(-C_{ik}/\tau_c)}{\sum_{l=1}^{K_t}\pi_l \exp(-C_{il}/\tau_c)}. \tag{39}$$

The overall coding cost additionally includes the disagreement penalty:

$$\mathcal{R}_i(\mathcal{C}_t) := -\tau_c \log \sum_{k=1}^{K_t} \pi_k \exp(-C_{ik}/\tau_c) + \gamma D_i. \tag{40}$$

This decomposition is algebraically equivalent to substituting $E_{ik} = C_{ik} + \gamma D_i$ into the original free-energy expression. The common factor induced by D_i cancels from a_{ik} but remains in $\mathcal{R}_i(\mathcal{C}_t)$. Hence, a_{ik} determines which known class best matches the node, while $\mathcal{R}_i(\mathcal{C}_t)$ determines whether the current class space can represent it.

The weight γ controls the contribution of cross-modal disagreement to unknown rejection. Since γD_i cancels from the class cross-entropy, γ is a fixed hyperparameter rather than a parameter learned by Eq. 7. It is fixed before constructing the calibration references and remains unchanged during calibration and testing. The distributions used to compute D_i are learned through C_{ik} and ω_i^r, but the disagreement term itself receives no direct gradient from the class cross-entropy.

Because D_i is also independent of the current class set, it cancels from cost reductions of the form $\mathcal{R}_i(\mathcal{C}_t) - \mathcal{R}_i(\mathcal{C}_t \cup \{p\})$. Thus, disagreement directly affects the initial known–unknown partition and influences class induction indirectly through the selected unknown nodes.

Known-reference calibration. Raw coding costs are not directly comparable across datasets or predicted classes. We therefore construct predicted-class reference distributions from a held-out set of trusted known nodes. For every calibration node j, let

$$\hat{c}_j := \arg\max_k a_{jk}, \qquad \mathcal{D}_K(c) := \left\{ \mathcal{R}_j(\mathcal{C}_t) \mid j \in \mathcal{V}^K_{\text{cal}}, \hat{c}_j = c \right\}. \tag{41}$$

For a test node i, let $\hat{c}_i = \arg\max_k a_{ik}$. MOVE computes its finite-sample upper-tail score as

$$p_i := \frac{1 + \sum_{s \in \mathcal{D}_K(\hat{c}_i)} \mathbb{I}[s \geq \mathcal{R}_i(\mathcal{C}_t)]}{1 + |\mathcal{D}_K(\hat{c}_i)|}, \quad b_i = -\log_2 p_i. \tag{42}$$

Using the same predicted-class partition for calibration and test nodes avoids conditioning the two sets on different class definitions. A small p_i indicates that the node has an unusually high coding cost relative to trusted known nodes receiving the same predicted class.

The tail-probability cutoffs α_U and α_K are fixed before test evaluation and satisfy $\alpha_U < \alpha_K$. MOVE then performs

$$\text{state}(i) := \begin{cases} \text{known}, & p_i > \alpha_K, \\ \text{pending}, & \alpha_U < p_i \leq \alpha_K, \\ \text{unknown}, & p_i \leq \alpha_U. \end{cases} \tag{43}$$

Here, α_U defines the high-confidence unknown tail, while α_K separates accepted known nodes from ambiguous pending nodes. Known nodes retain their existing class assignments. Unknown nodes seed new-class discovery. Pending nodes never initiate new classes and are used only as supporting context when they are sufficiently close to an unknown group.

B.3 Evidence-Grounded Class Induction

Fused evidence distribution. For each unknown node, the three modality-specific evidence distributions are fused with a reliability-weighted product of experts:

$$q_i(\mathbf{z}) = \frac{1}{Z_i} \prod_{r \in \{v,t,g\}} [q_i^r(\mathbf{z})]^{\omega_i^r}. \tag{44}$$

Because all experts are Gaussian, the fused distribution also has a closed-form Gaussian representation:

$$\mathbf{\Lambda}_i = \sum_r \omega_i^r \mathbf{\Lambda}_i^r, \quad \boldsymbol{\mu}_i = \mathbf{\Lambda}_i^{-1} \sum_r \omega_i^r \mathbf{\Lambda}_i^r \boldsymbol{\mu}_i^r, \quad \mathbf{\Lambda}_i^r = (\mathbf{\Sigma}_i^r)^{-1}. \tag{45}$$

This fused distribution summarizes the node-level multimodal evidence used for grouping and subsequent class induction.

Unknown-node grouping. MOVE does not assume that the number of unseen classes is known in advance. We therefore group unknown nodes according to the Bhattacharyya distance between their fused evidence distributions:

$$\{S_g\}_{g=1}^G = \text{AggCluster}\left(\{q_i : \text{state}(i) = \text{unknown}\}, D_{\text{B}}, \tau_{\text{grp}}\right). \tag{46}$$

We use agglomerative grouping with a distance threshold τ_{grp}, so the number of discovered groups G is determined by the observations rather than supplied as an oracle. Nodes within the same group therefore have similar joint visual, textual, and structural evidence. Groups that contain too few high-confidence unknown nodes are deferred rather than immediately triggering class generation.

Proposal–validation split. Each sufficiently supported group is randomly divided into two disjoint subsets:

$$S_g = S_g^P \cup S_g^V, \quad S_g^P \cap S_g^V = \emptyset. \tag{47}$$

The proposal subset S_g^P is used to generate and ground candidate classes. The validation subset S_g^V is never exposed to the MLLM or to candidate construction and is reserved exclusively for the selective-expansion stage. This separation is essential because evaluating a candidate on the same observations used to generate it would systematically favor overfitted or overly specific classes.

Multimodal evidence package. For every proposal group, MOVE forms an evidence package $\mathcal{E}_g$ from representative nodes in S_g^P. The package contains the corresponding visual inputs, textual attributes, and compact graph-context summaries derived from their local neighborhoods. The graph summary includes structural and neighbor information available in the underlying dataset, while avoiding unavailable ground-truth labels of unknown nodes. The evidence package therefore exposes the MLLM to the same three information sources used by MOVE rather than relying on natural-language node descriptions alone.

Candidate class proposal. The multimodal LLM receives $\mathcal{E}_g$ and returns a set of M concise candidate class descriptions:

$$\mathcal{H}_g = \{h_{g,m}\}_{m=1}^{M}. \tag{48}$$

These descriptions are treated strictly as proposals. The MLLM does not decide which candidate is accepted, nor does linguistic plausibility alone constitute evidence for expanding the class space.

Candidate encoding. Each natural-language candidate description is mapped back to the same semantic space used by the node representations. Specifically, we encode the candidate with the textual semantic encoder and the corresponding distribution head:

$$\mathbf{e}_{g,m}^h = f_t(h_{g,m}), \quad (\boldsymbol{\mu}_{g,m}^h, \boldsymbol{\sigma}_{g,m}^h) = \phi_t(\mathbf{e}_{g,m}^h), \quad q_{g,m}^h = \mathcal{N}\left(\boldsymbol{\mu}_{g,m}^h, \boldsymbol{\Sigma}_{g,m}^h\right). \tag{49}$$

The resulting $q_{g,m}^h$ is the *candidate prior*: it expresses where the MLLM-generated semantic description lies in the shared latent space before observing the proposal-group evidence. It is therefore not yet the final class representation.

Evidence grounding. MOVE combines the language-derived candidate prior with the multimodal evidence of the proposal nodes. Let $\boldsymbol{\Lambda}_{g,m}^h = (\boldsymbol{\Sigma}_{g,m}^h)^{-1}$. We assign each proposal observation the normalized confidence weight

$$w_i = \frac{\exp[-\mathcal{H}(q_i)/\tau_\nu]}{\sum_{j\in S_g^P} \exp[-\mathcal{H}(q_j)/\tau_\nu]}, \qquad \sum_{i\in S_g^P} w_i = 1. \tag{50}$$

To represent the population distribution of the candidate class, rather than the posterior uncertainty of its sample mean, we construct the moment-matched group distribution

$$\begin{aligned} \bar{\boldsymbol{\mu}}_g &= \sum_{i\in S_g^P} w_i \boldsymbol{\mu}_i, \\ \bar{\boldsymbol{\Sigma}}_g &= \sum_{i\in S_g^P} w_i \left[\boldsymbol{\Sigma}_i + (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_g)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_g)^\top\right] + \varepsilon_\Sigma \mathbf{I}, \\ \bar{\boldsymbol{\Lambda}}_g &= \bar{\boldsymbol{\Sigma}}_g^{-1}. \end{aligned} \tag{51}$$

The first term preserves within-node uncertainty, while the outer-product term captures between-node population scatter. The ridge $\varepsilon_\Sigma \mathbf{I}$ ensures invertibility.

The candidate prior and the pooled evidence distribution are combined by

$$\begin{aligned} \boldsymbol{\Lambda}_{g,m}^* &= \frac{\beta \boldsymbol{\Lambda}_{g,m}^h + \bar{\boldsymbol{\Lambda}}_g}{\beta + 1}, \\ \boldsymbol{\mu}_{g,m}^* &= \left(\beta \boldsymbol{\Lambda}_{g,m}^h + \bar{\boldsymbol{\Lambda}}_g\right)^{-1} \left(\beta \boldsymbol{\Lambda}_{g,m}^h \boldsymbol{\mu}_{g,m}^h + \bar{\boldsymbol{\Lambda}}_g \bar{\boldsymbol{\mu}}_g\right). \end{aligned} \tag{52}$$

The grounded candidate is

$$p_{g,m}^* = \mathcal{N}\left(\boldsymbol{\mu}_{g,m}^*, (\boldsymbol{\Lambda}_{g,m}^*)^{-1}\right). \tag{53}$$

Because the evidence weights sum to one and the pooled precision is normalized by $\beta + 1$, the candidate covariance does not shrink merely because the proposal group contains more observations.

Proposal-set support. Before entering the expansion stage, each grounded candidate is evaluated on the proposal subset:

$$u_{g,m} = \frac{1}{|S_g^P|} \sum_{i \in S_g^P} \left[\mathcal{R}_i(\mathcal{C}_t) - \mathcal{R}_i(\mathcal{C}_t \cup \{p_{g,m}^*\}) \right], \quad \rho_{g,m} = \frac{\exp(u_{g,m}/\tau_h)}{\sum_j \exp(u_{g,j}/\tau_h)}. \tag{54}$$

Here, $u_{g,m}$ measures how much the candidate reduces assignment cost on the observations that proposed it, while $\rho_{g,m}$ provides a normalized support ranking among candidates from the same group. Importantly, $\rho_{g,m}$ is not an admission probability: final acceptance depends only on independent validation and the selective-expansion criterion described below.

B.4 Selective Class Expansion

Provisional class prior. Evaluating $\mathcal{R}_i(\mathcal{C}_t \cup \{p_{g,m}^*\})$ requires a provisional prior for the candidate class. To avoid favoring candidates solely through an arbitrarily large mixture weight, we initialize the candidate prior from the size of its supporting proposal group:

$$\pi_g^{\text{tmp}} = \frac{|S_g^P|}{N_K + |S_g^P|}, \quad \pi_k^{\text{tmp}} = (1 - \pi_g^{\text{tmp}})\pi_k, \tag{55}$$

where N_K denotes the effective number of known reference observations. All candidates generated from the same group therefore use the same provisional prior, ensuring that their comparison depends on their representations rather than manually assigned prior mass.

Held-out generalization gain. The validation subset S_g^V is independent of candidate generation. For each validation node, we measure the reduction in coding cost after tentatively introducing candidate $p_{g,m}^*$:

$$\Delta_{i,g,m} = \mathcal{R}_i(\mathcal{C}_t) - \mathcal{R}_i(\mathcal{C}_t \cup \{p_{g,m}^*\}). \tag{56}$$

Let $\widehat{\sigma}_{g,m}$ denote the sample standard deviation of $\{\Delta_{i,g,m}\}_{i \in S_g^V}$. MOVE uses a one-sided lower confidence bound:

$$G_{g,m}^{\text{LCB}} = \frac{1}{|S_g^V|} \sum_{i \in S_g^V} \Delta_{i,g,m} - z_{1-\delta_{\text{conf}}} \frac{\widehat{\sigma}_{g,m}}{\sqrt{|S_g^V|}}. \tag{57}$$

A candidate therefore receives a large generalization score only when it consistently improves nodes that were not used to create or ground it. Here $\delta_{\text{conf}} \in (0, 1)$ is the one-sided tail probability and $z_{1-\delta_{\text{conf}}}$ is the corresponding standard-normal quantile; this symbol is distinct from the semantic-complexity weight α defined below.

Redundancy and complexity control. A candidate that closely overlaps an existing class should not be added merely because it reduces local assignment cost. We therefore define a soft redundancy score

$$R_{g,m}^{\text{red}} = \sum_{k=1}^{K_t} \pi_k \exp\left(-\frac{D_{\text{B}}(p_{g,m}^*, p_k)}{\tau_\Omega} \right), \tag{58}$$

which approaches one when the candidate is close to a high-probability existing class and decreases as it becomes semantically distinct. We further account for the effective complexity of introducing a new diagonal-Gaussian class:

$$C_{g,m}^{\text{cmp}} = \frac{d_{\text{eff}}}{2|S_g^P|} \log(1 + |S_g^P|), \quad \Omega_{g,m} = R_{g,m}^{\text{red}} + \lambda_{\text{cmp}} C_{g,m}^{\text{cmp}}. \tag{59}$$

For a d-dimensional diagonal Gaussian, $d_{\text{eff}} = 2d$ when the mean and variance parameters are counted. The normalization by $|S_g^P|$ prevents a candidate supported by many observations from being penalized identically to one supported by only a few samples.

Operational hyperparameters in the robustness analysis. For reproducibility, we formally define the two implementation-level hyperparameters varied in Fig. 4. Let $\mathbf{z}_{g,m}$ be the normalized embedding of candidate $h_{g,m}$ and let $\{\mathbf{c}_k\}_{k=1}^{K_t}$ be the normalized active-class prototypes. We compute the candidate residual and a dictionary-calibrated reference residual as

$$r_{g,m} = \min_k \|\mathbf{z}_{g,m} - \mathbf{c}_k\|_2^2, \qquad r_{\mathrm{ref}} = \operatorname*{median}_k \min_{\ell \neq k} \|\mathbf{c}_k - \mathbf{c}_\ell\|_2^2, \tag{60}$$

and define the conditional semantic code length and its normalized multiplier by

$$\begin{aligned} C_{g,m}^{\mathrm{sem}} &= \log_2\left(1 + \frac{r_{g,m}}{r_{\mathrm{ref}} + \varepsilon_{\mathrm{cmp}}}\right), \\ \omega_{g,m}^{\mathrm{sem}}(\alpha) &= \mathrm{clip}\left(\frac{1 + \alpha C_{g,m}^{\mathrm{sem}}}{1+\alpha}, \omega_{\min}, \omega_{\max}\right), \qquad \alpha \geq 0. \end{aligned} \tag{61}$$

The implementation-level knowledge-complexity cost is $L_{g,m}^{\mathrm{cmp,op}} = \lambda \omega_{g,m}^{\mathrm{sem}}(\alpha)$, where $\lambda \geq 0$ is the knowledge-complexity penalty shown on the horizontal axis of Fig. 4(a–b), and α controls how strongly candidate-specific semantic distance modulates that penalty. This cost replaces the parametric complexity contribution $\lambda_\Omega \lambda_{\mathrm{cmp}} C_{g,m}^{\mathrm{cmp}}$ in Eq. 65, while the redundancy and interference terms remain unchanged. The operational admission score used in the robustness analysis is therefore

$$\mathcal{A}_{g,m}^{\mathrm{op}} = G_{g,m}^{\mathrm{LCB}} - \lambda_\Omega R_{g,m}^{\mathrm{red}} - \lambda \omega_{g,m}^{\mathrm{sem}}(\alpha) - \lambda_\Xi \Xi_{g,m}. \tag{62}$$

The candidate is admitted when $\mathcal{A}_{g,m}^{\mathrm{op}} \geq \tau_A$, with $\tau_A = 0$ by default. Figures 4(a–b) vary only λ and α while holding the redundancy, interference, and validation terms fixed. Thus, $\alpha = 0$ recovers a constant unit complexity, and clipping prevents a single semantic residual from dominating the admission score.

An *unknown anchor* is a high-confidence unknown node, i.e., a node with $p_i \leq \alpha_U$ in Eq. 43. For a proposal group S_g, define its unknown-anchor fraction and evidence gate as

$$f_g^U = \frac{1}{|S_g|} \sum_{i \in S_g} \mathbf{1}\{p_i \leq \alpha_U\}, \qquad \mathrm{gate}_\eta(S_g) = \mathbf{1}\{f_g^U \geq \eta\}, \quad 0 \leq \eta \leq 1. \tag{63}$$

Only groups passing this gate are eligible to trigger candidate admission; otherwise they are deferred until further evidence arrives. The robustness analysis varies $\eta =$ `min_unknown_anchor_fraction` and uses $\eta = 0.5$ by default. Thus, increasing η requires a larger proportion of high-confidence unknown evidence in each proposal group.

Interference with existing classes. A valid new class should also avoid destabilizing decisions for already known observations. Let $\mathcal{B}_{\mathrm{old}}$ be a class-balanced buffer of trusted known nodes. Denote their class-assignment vectors before and after adding candidate $p_{g,m}^*$ by $\mathbf{a}_i^- \in \mathbb{R}^{K_t}$ and $\mathbf{a}_{i,g,m}^+ \in \mathbb{R}^{K_t+1}$, respectively. We pad the original assignment by one zero entry and define

$$\Xi_{g,m} = \frac{1}{|\mathcal{B}_{\mathrm{old}}|} \sum_{i \in \mathcal{B}_{\mathrm{old}}} D_{\mathrm{JS}}\left([\mathbf{a}_i^-, 0] \| \mathbf{a}_{i,g,m}^+\right). \tag{64}$$

A large $\Xi_{g,m}$ indicates that introducing the candidate substantially changes the predictions of known nodes and therefore risks damaging previously learned class boundaries.

Final admission decision. MOVE combines held-out improvement, redundancy/complexity, and old-class interference:

$$\mathcal{A}_{g,m} = G_{g,m}^{\mathrm{LCB}} - \lambda_\Omega \Omega_{g,m} - \lambda_\Xi \Xi_{g,m}, \quad m^* = \arg\max_m \mathcal{A}_{g,m}. \tag{65}$$

The selected candidate is accepted only if its net admission score exceeds a conservative threshold τ_A:

$$\mathcal{C}_{t+1} = \begin{cases} \mathcal{C}_t \cup \{(h_{g,m^*}, \pi_{g,m^*}, p_{g,m^*}^*)\}, & \mathcal{A}_{g,m^*} \geq \tau_A, \\ \mathcal{C}_t, & \mathcal{A}_{g,m^*} < \tau_A. \end{cases} \tag{66}$$

We use $\tau_A = 0$ as the default decision rule, recovering the positive-net-gain criterion in the main text. The semantic description h_{g,m^*} is retained as the generated class label used for unknown-class annotation, while p_{g,m^*}^* becomes the distributional representation used for subsequent class assignment. The new class prior π_{g,m^*} is initialized from its supporting group size and all class priors are renormalized after admission.

Sequential expansion. When several unknown groups are discovered at the same stage, candidates are processed in descending order of their admission scores. Once a class is accepted, $\mathcal{C}_t$ is immediately updated before evaluating subsequent groups. Consequently, later candidates are also compared against classes accepted earlier in the same expansion stage, reducing the possibility that two independently discovered groups introduce duplicate classes.

In a multi-stage deployment, the expanded class space $\mathcal{C}_{t+1}$, including the admitted class representations and renormalized priors, is used as the reference space for the next arrival stage. MOVE then reapplies the known/pending/unknown decision to newly arrived nodes and repeats candidate induction and selective admission only for newly unresolved groups, yielding $\mathcal{C}_{t+2}$ when additional reliable classes are accepted. Since each new candidate is evaluated against the accumulated class space, classes admitted in earlier stages also participate in the redundancy check.

B.5 Training and Deployment Procedure

Known-class training. The modality projections, graph encoder, distribution heads, and known-class representations are optimized only with labeled known-class training nodes. The coding loss in Eq. 7 acts on the class-dependent compatibility C_{ik}. The disagreement D_i is computed from the learned modality distributions but receives no direct gradient from this loss, and γ remains fixed throughout training. No unknown-class labels are used to optimize the coding model.

Open-world deployment. Before deployment, γ, α_U, and α_K are fixed, and the reference distributions $\mathcal{D}_K(c)$ are constructed from the held-out known-class calibration set. MOVE then applies the trained coding model to unlabeled nodes and performs the three-way known/pending/unknown decision. Only high-confidence unknown nodes seed new-class discovery. Candidate generation and grounding use the proposal subsets, while admission uses the disjoint validation subsets and the trusted old-class buffer. Unknown ground-truth labels are never accessed during proposal, grounding, or expansion.

Information flow across stages. The complete execution path is therefore

$$(\mathbf{x}_i^v, \mathbf{x}_i^t, \mathbf{A}) \to (\mathbf{e}_i^v, \mathbf{e}_i^t, \mathbf{e}_i^g) \to (q_i^v, q_i^t, q_i^g) \to \mathrm{state}(i) \to S_g \to (S_g^P, S_g^V) \to \mathcal{H}_g \to p_{g,m}^* \to \mathcal{A}_{g,m} \to \mathcal{C}_{t+1}.$$

This data flow corresponds directly to Fig. 3: multimodal graph encoding produces visual, textual, and topology-conditioned representations; uncertainty-aware class coding identifies unresolved nodes; evidence-grounded induction converts grouped unknown observations into grounded class candidates; and selective expansion determines which candidates are sufficiently general, non-redundant, and non-interfering to enter the evolving class space.

C Theoretical Analysis

We analyze two conditional properties of MOVE's coding and admission objectives. The first concerns a sample-mean admission rule for a fixed candidate under independent validation sampling and bounded coding gains. The second bounds coding-cost changes when modality terms are added with fixed reliability weights.

C.1 Consistency of Fixed-Candidate Mean-Based Admission

Fix the concept codebook $\mathcal{K}$, a candidate concept p, the coding function and its parameters, and the unresolved-observation distribution $\mathcal{D}_u$ before drawing validation samples. Candidate generation and these choices must not use the validation set. We define the population compression gain of p as

$$\Delta(p) = \mathbb{E}_{x \sim \mathcal{D}_u} \left[R(x; \mathcal{K}) - R(x; \mathcal{K} \cup \{p\}) \right], \tag{67}$$

where $\mathcal{D}_u$ denotes the distribution of unresolved observations and $R(\cdot)$ is the conditional coding cost defined by MOVE.

The ideal admission rule is to accept p when its expected explanatory gain exceeds the additional knowledge complexity:

$$A^*(p) = \Delta(p) - \lambda_\Omega \Omega(p) - \lambda_\Xi \Xi(p) > 0. \tag{68}$$

Hold λ_Ω, λ_Ξ, $\Omega(p)$, and $\Xi(p)$ fixed independently of the validation set. For a fixed sample size $n = |\mathcal{S}^V|$, consider the sample-mean admission rule that accepts p if and only if the following score is positive:

$$\hat{A}(p) = \frac{1}{|\mathcal{S}^V|} \sum_{x_i \in \mathcal{S}^V} (R(x_i; \mathcal{K}) - R(x_i; \mathcal{K} \cup \{p\})) - \lambda_\Omega \Omega(p) - \lambda_\Xi \Xi(p). \tag{69}$$

Theorem 1 (Fixed-Candidate Mean Admission). Under the fixed setup above, let the validation samples be independent and identically distributed according to $\mathcal{D}_u$. Assume that the coding gain variable

$$Z_p(x) = R(x; \mathcal{K}) - R(x; \mathcal{K} \cup \{p\})$$

satisfies $|Z_p(x)| \le M$ almost surely for a constant $M > 0$. For this fixed candidate p and a fixed $\epsilon > 0$, suppose that the population admission margin satisfies

$$|A^*(p)| > \epsilon, \tag{70}$$

Then the sign-error probability of this sample-mean rule satisfies

$$\Pr\left(\operatorname{sign}(\hat{A}(p)) \neq \operatorname{sign}(A^*(p))\right) \le 2 \exp\left(-\frac{|\mathcal{S}^V|\epsilon^2}{2M^2}\right). \tag{71}$$

Consequently, for this fixed setup with $A^*(p) \neq 0$, as the sample size increases,

$$\lim_{|\mathcal{S}^V| \to \infty} \Pr\left(\operatorname{sign}(\hat{A}(p)) = \operatorname{sign}(A^*(p))\right) = 1. \tag{72}$$

Proof. With the candidate, coding function, and penalties fixed, independent samples from $\mathcal{D}_u$ give an unbiased estimate of the population admission score:

$$\mathbb{E}[\hat{A}(p)] = A^*(p). \tag{73}$$

The fixed penalty terms cancel in $\hat{A}(p) - A^*(p)$. A sign error requires $|\hat{A}(p) - A^*(p)| \ge |A^*(p)| > \epsilon$, and hence

$$|\hat{A}(p) - A^*(p)| > \epsilon. \tag{74}$$

Applying Hoeffding's inequality to the bounded random variable $Z_p(x) \in [-M, M]$, whose interval length is $2M$, gives

$$\Pr\left(|\hat{A}(p) - A^*(p)| > \epsilon\right) \le 2 \exp\left(-\frac{|\mathcal{S}^V|\epsilon^2}{2M^2}\right). \tag{75}$$

For any fixed nonzero margin, choose $0 < \epsilon < |A^*(p)|$. The bound then vanishes as the sample size increases, proving pointwise consistency of the sample-mean rule for this candidate. □

This result concerns agreement with the population penalized-gain rule, rather than semantic correctness of the candidate concept. It does not provide a uniform guarantee over adaptively selected candidates or repeated reuse of validation data, nor does it cover admission statistics other than the sample mean.

C.2 Coding-Cost Separation with Fixed Weights

MOVE uses uncertainty-aware multimodal coding to identify unknown observations. We analyze coding-cost increments under fixed reliability weights and explicit modality-mismatch conditions.

Let p_k denote an existing concept and q_i^r denote the evidence distribution of node i under modality $r \in \{v, t, g\}$. The concept-conditioned energy is defined as

$$E_{ik} = \sum_r \omega_i^r D(q_i^r \| p_k) + \gamma \sum_{r<s} \omega_i^r \omega_i^s D_B(q_i^r, q_i^s). \tag{76}$$

The corresponding coding cost is

$$R_i(\mathcal{K}) = -\tau_c \log \sum_k \pi_k \exp(-E_{ik}/\tau_c). \tag{77}$$

We consider two types of observations. An explainable observation satisfies that there exists an existing concept p_k such that all evidence sources remain close to this concept. In contrast, an unexplained observation has a non-zero multimodal mismatch margin:

$$\min_k E_{ik} \geq \delta, \tag{78}$$

where $\delta > 0$ measures the energy gap relative to the current codebook. These descriptions alone do not imply the incremental conditions below; the separation result requires those additional conditions.

Theorem 2 (Fixed-Weight Conditional Separation). Condition on a fixed codebook, evidence distributions, and nonnegative reliability weights $\{\omega_i^r\}$. Fix $\gamma \geq 0$, $\tau_c > 0$, and class priors $\pi_k \geq 0$ with $\sum_k \pi_k = 1$, and assume finite energies. Let $\mathcal{M}$ be a fixed set of m modalities. Define $E_{ik}^{(m)}$ by restricting the energy sums to $\mathcal{M}$, and $E_{ik}^{(m+1)}$ by extending them to $\mathcal{M} \cup \{r'\}$. All weights and other quantities remain unchanged in this comparison; in particular, reliability weights are neither re-estimated nor renormalized. For nonnegative divergences, the added modality contributes

$$B_{ik}^{r'} = \omega_i^{r'} D(q_i^{r'} \| p_k) + \gamma \sum_{r \in \mathcal{M}} \omega_i^r \omega_i^{r'} D_B(q_i^r, q_i^{r'}),$$

where D_B is symmetric, and thus

$$B_{ik}^{r'} = E_{ik}^{(m+1)} - E_{ik}^{(m)} \geq 0 \tag{79}$$

to class k. Define the assignment distribution based on the first m terms,

$$a_{ik}^{(m)} = \frac{\pi_k \exp(-E_{ik}^{(m)}/\tau_c)}{\sum_\ell \pi_\ell \exp(-E_{i\ell}^{(m)}/\tau_c)}. \tag{80}$$

For an unexplained observation, suppose that $B_{ik}^{r'} \geq \delta_{r'} > 0$ on a set of classes carrying mass at least $\rho \in (0, 1]$ under $a_i^{(m)}$. Then

$$R_i^{(m+1)}(\mathcal{K}) - R_i^{(m)}(\mathcal{K}) \geq -\tau_c \log\Big[1 - \rho\left(1 - e^{-\delta_{r'}/\tau_c}\right)\Big]. \tag{81}$$

In particular, if every existing class has incremental mismatch at least $\delta_{r'}$, then the increase is at least $\delta_{r'}$. Conversely, if an explainable observation satisfies $0 \leq B_{ik}^{r'} \leq \epsilon_{r'}$ for every class with nonzero assignment mass, then

$$0 \leq R_i^{(m+1)}(\mathcal{K}) - R_i^{(m)}(\mathcal{K}) \leq \epsilon_{r'}. \tag{82}$$

For a pair of observations satisfying the respective conditions, the change in their unexplained–explainable coding-cost gap is bounded below by Eq. 81 minus $\epsilon_{r'}$. The gap increases whenever this lower bound is positive. The same increment bounds hold in expectation if the conditions hold almost surely with common constants and the increments are integrable.

Proof. Using Eq. 80 and $E_{ik}^{(m+1)} = E_{ik}^{(m)} + B_{ik}^{r'}$, the change in coding cost has the exact form

$$R_i^{(m+1)} - R_i^{(m)} = -\tau_c \log \frac{\sum_k \pi_k \exp[-(E_{ik}^{(m)} + B_{ik}^{r'})/\tau_c]}{\sum_k \pi_k \exp(-E_{ik}^{(m)}/\tau_c)} \tag{83}$$

$$= -\tau_c \log \sum_k a_{ik}^{(m)} \exp(-B_{ik}^{r'}/\tau_c). \tag{84}$$

Because $B_{ik}^{r'} \geq 0$, the weighted exponential sum in Eq. 84 is at most one, proving the nonnegative lower bound. Let $\mathcal{I}$ be the set of classes for which $B_{ik}^{r'} \geq \delta_{r'}$ and let $\sum_{k \in \mathcal{I}} a_{ik}^{(m)} \geq \rho$. Then

$$\sum_k a_{ik}^{(m)} e^{-B_{ik}^{r'}/\tau_c} \leq (1 - \rho) + \rho e^{-\delta_{r'}/\tau_c}. \tag{85}$$

Since $-\log(\cdot)$ is decreasing, substituting this upper bound into Eq. 84 proves Eq. 81. Setting $\rho = 1$ reduces its right-hand side to $\delta_{r'}$.

For an explainable observation, $B_{ik}^{r'} \leq \epsilon_{r'}$ implies $e^{-B_{ik}^{r'}/\tau_c} \geq e^{-\epsilon_{r'}/\tau_c}$ for every class with nonzero assignment mass. Therefore,

$$\sum_k a_{ik}^{(m)} e^{-B_{ik}^{r'}/\tau_c} \geq e^{-\epsilon_{r'}/\tau_c}, \tag{86}$$

and Eq. 84 gives the upper bound $R_i^{(m+1)} - R_i^{(m)} \leq \epsilon_{r'}$. Subtracting this upper bound from the unexplained-observation lower bound yields the claimed conditional separation result. Taking expectations preserves these inequalities under the stated integrability assumption. □

The weights may depend on each observation, but are held fixed when comparing its two coding costs. The theorem does not cover recomputing or renormalizing weights, updating concepts, or refitting the model after adding a modality.

D Experiment in Details

Datasets. We evaluate MOVE on four multimodal attributed graph datasets, including Toys, Grocery, Movies, and Reddit-S Yan et al. (2025). These datasets represent different real-world scenarios with heterogeneous node attributes and relational structures. Each node is associated with multimodal information, including textual attributes, visual content, and graph connections. Toys and Grocery represent product-centered graphs where nodes contain item descriptions, images, and user interaction structures. Movies represents a multimedia recommendation graph containing textual metadata, visual information, and relational links among entities. Reddit-S represents a social graph where nodes are associated with semantic attributes and network structures. These datasets provide diverse evaluation environments for testing open-world class recognition and expansion.

Open-world Protocol. Following the open-world graph learning setting, we divide the original class space into known and unknown categories. Specifically, 50% of the classes are randomly selected as known classes, whose nodes are available during training, while the remaining 50% are treated as unknown classes that only appear during evaluation. The model is trained exclusively with known-class nodes and is required to recognize known nodes, identify unknown nodes, infer their potential classes, and determine whether new classes should be incorporated into the evolving class space. For downstream evaluation, we compare three settings: the original graph without unknown-class annotation, the graph annotated by MOVE, and the graph annotated using ground-truth unknown labels.

Repeated Runs and Uncertainty. All reported results are averaged over five independently seeded runs. Tables with uncertainty terms report the mean and standard deviation across these runs. Because of the width of Tables 2 and 5, these two tables display the five-seed means only; the omission of standard deviations is a presentation choice rather than a single-run evaluation.

Baselines. We compare MOVE with representative methods from three categories. The first category contains OOD detection methods, including TopoOOD Bao et al. (2024) and GNN-Safe Wu et al. (2023), which focus on identifying nodes outside the training distribution but do not explicitly generate new class descriptions. The second category contains structural open-world graph learning methods, including ORAL Jin et al. (2024) and OpenWGL Wu et al. (2020), which identify or organize unknown nodes. For semantic label quality evaluation, their outputs are converted into natural-language labels using the common adaptation described below. The third category contains multimodal open-world methods, including OGA Wen et al. (2025) and AHM Constantinou et al. (2026), which leverage multimodal or LLM-based information for open-world recognition. We additionally compare with recent graph-based open-world methods, including CFC Ma et al. (2026) and TopoNC Xi et al. (2026), to evaluate the effectiveness of topology-aware unknown recognition. Methods without native class annotation or expansion capabilities are evaluated only on the tasks supported by their original designs.

Semantic-label adaptation. Since OpenWGL and ORAL do not natively generate natural-language class labels, we adapt them to the UCA evaluation as follows. We retain the unknown-node predictions produced by each method, group the predicted unknown nodes according to graph connectivity, and summarize the representative multimodal evidence of each group into a concise semantic label

using LLM. This adaptation is used only to provide the natural-language outputs required by the semantic label quality metrics and does not modify the original unknown-node recognition results.

Evaluation Tasks. We evaluate MOVE from three perspectives following the complete open-world pipeline.

Unknown-Class Rejection (UCR). The first task evaluates whether the model can distinguish unknown nodes from known classes. We report known-class recognition accuracy and unknown-node rejection performance. Accuracy measures whether known categories are preserved after open-world learning, while Coverage and Precision evaluate the ability to identify unknown nodes.

Unknown-Class Annotation (UCA). The second task evaluates whether the discovered classes accurately describe previously unseen concepts. We measure the semantic similarity between generated class descriptions and ground-truth unknown labels using G-to-U similarity, where higher values indicate better alignment with true unknown categories. We also report K-to-G similarity, which measures the similarity between generated classes and known classes, where lower values indicate fewer redundant expansions.

Downstream Graph Learning. Finally, we evaluate whether the discovered classes provide useful information for downstream graph learning. We conduct node classification using a GCN Kipf & Welling (2016) backbone and report classification accuracy. We also evaluate zero-shot modality matching to measure whether the generated classes preserve multimodal semantic information. For node classification, the graph without unknown-class annotation and the graph with ground-truth unknown labels are used as lower and upper bounds, respectively.

Metric Definitions. Let $\mathcal{Z}^G$, $\mathcal{Z}^U$, and $\mathcal{Z}^K$ denote the unit-normalized semantic prototypes of generated, held-out unknown, and known classes, respectively. All prototypes lie in the shared CLIP ViT-L/14 semantic space. We compute

$$\text{G-to-U} = \frac{1}{|\mathcal{Z}^G|} \sum_{\mathbf{z}_g \in \mathcal{Z}^G} \max_{\mathbf{z}_u \in \mathcal{Z}^U} \mathbf{z}_g^\top \mathbf{z}_u, \qquad \text{K-to-G} = \frac{1}{|\mathcal{Z}^K|} \sum_{\mathbf{z}_k \in \mathcal{Z}^K} \max_{\mathbf{z}_g \in \mathcal{Z}^G} \mathbf{z}_k^\top \mathbf{z}_g. \tag{87}$$

Thus, G-to-U measures how closely each generated concept matches its nearest unknown-class prototype, whereas K-to-G measures the nearest generated-concept similarity for each known class. Higher G-to-U and lower K-to-G indicate accurate unknown semantics with less overlap with the known label space.

For zero-shot modality matching, each held-out unknown test image is a query and all generated concept descriptions form the candidate pool. Queries and candidate descriptions are encoded by the CLIP ViT-L/14 image and text encoders, ranked by cosine similarity, and evaluated using Recall@5. A query is counted as correct when at least one of its five highest-ranked generated concepts is associated with the query's ground-truth unknown class through that concept's support nodes. Ground-truth labels are used only to score relevance and never to construct or rank candidates. The MM values in Table 4 are Recall@5 percentages:

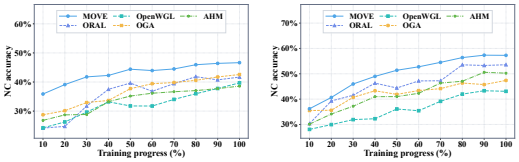
$$\text{MM-Recall@5} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}\left[y_q \in \bigcup_{j \in \text{Top5}(q)} \mathcal{Y}_j \right], \tag{88}$$

where y_q is the unknown-class label of query q and $\mathcal{Y}_j$ is the set of unknown classes represented by the support nodes of generated concept j.

Ablation Variants. We evaluate five functional components of MOVE. The corresponding ablation variants are defined as follows.

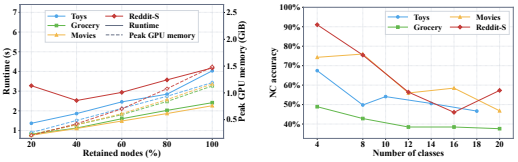
Cross-Modal Disagreement Modeling (CDM). CDM captures inconsistency among modality-specific evidence through the pairwise disagreement term in Eq. 5. The *w/o CDM* variant removes this term, so that class coding depends only on the compatibility between individual modalities and existing classes.

Held-out Candidate Validation (HCV). HCV evaluates grounded candidate classes on the independent validation subset S_g^V. The *w/o HCV* variant selects candidates using only the proposal subset S_g^P, without testing their generalization to held-out observations.



(a) Toys. (b) Reddit-S.

Figure 7: **Convergence comparison.** Post-annotation NC accuracy over normalized upstream training progress.



(a) Graph size. (b) Category size.

Figure 8: **Scalability.** (a) Runtime and peak GPU memory versus graph size. (b) Post-annotation NC accuracy versus number of classes.

Evidence Grounding (EG). EG grounds the language-derived candidate prior in the visual, textual, and graph-context evidence of the proposal group. The *w/o EG* variant directly uses the original CLIP-encoded language prior while leaving the subsequent validation and class-space update unchanged.

Graph-Context Encoding (GCE). GCE incorporates topology-conditioned representations into multimodal class coding and candidate induction. The *w/o GCE* variant removes graph-context evidence and retains only visual and textual information.

Class-Space Update (CSU). CSU incorporates admitted candidate classes into the evolving class space $\mathcal{C}_{t+1}$. In the *w/o CSU* variant, generated classes may be used to annotate the corresponding unknown nodes but are not incorporated into the class space for subsequent learning.

Additional Convergence and Scalability Results.

Convergence Analysis. Fig. 7 reports post-annotation NC accuracy over normalized upstream training progress on Toys and Reddit-S. MOVE achieves the highest accuracy at every evaluated checkpoint and improves overall as training progresses. In contrast, the baseline curves exhibit larger fluctuations or converge to lower accuracy. These results indicate that MOVE maintains a stable performance advantage throughout training.

Scalability Analysis. Fig. 8(a) shows that runtime and peak GPU memory increase moderately with the retained-node ratio, remaining below 4.2 seconds and 1.5 GiB at full scale. Fig. 8(b) shows that NC accuracy generally decreases as the number of classes increases, reflecting the greater difficulty of a larger open-world label space. Nevertheless, MOVE retains 37.60%–57.27% accuracy under the complete class configurations. These results demonstrate manageable computational costs across the evaluated graph and category scales.