GOMA: TOWARD STRUCTURE-DRIVEN MULTIMODAL ALIGNMENT FROM A GRAPH SIGNAL SMOOTHING PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal retrieval uses images, text, and object relationships to answer different questions about the same collection. A query may seek an object's paired description, another object in the same category, or an object connected by an observed relationship. These goals rely on different notions of relevance. Paired matching requires object-specific distinctions, whereas cross-object retrieval benefits from relational agreement. Existing methods learn strong cross-modal correspondence or propagate information over a graph, but a shared output regularized toward neighbors can weaken identity distinctions. Moreover, gains from graph regularization can diminish after uniform graph propagation. We introduce **G**raph-**O**ptimized **M**ultimodal **A**lignment (GOMA), which assigns these roles to two connected embeddings. Each modality produces a content embedding directly supervised for paired identity and a semantic embedding jointly trained with cross-modal pairs and observed relationships. For complete records, both embeddings form an initial fused representation, semantic agreement sets positive weights on observed edges, and restart graph propagation reinjects this initial signal. This design lets a jointly trained model support single-modality and dual-attribute retrieval through a task-specific readout. Across six datasets and four tasks, GOMA achieves state-of-the-art performance on all 14 primary measures against 14 external methods. Controlled comparisons further show how separate supervision, graph regularization, and semantic-guided graph propagation shape the final representation and align the learned signal with each retrieval target.

1 INTRODUCTION

An image can initiate different searches. A user may want its paired description, another product with the same purpose, or an object linked through previous interactions. Images and text describe object attributes, while relationships such as co-purchases or post interactions organize the collection (Radford et al., 2021; Tschannen et al., 2025; Wei et al., 2024; Jiang et al., 2025; Yan et al., 2024; Wan et al., 2026). Because each search defines a different correct answer, multimodal alignment must expose the information required by its retrieval target.

Fig. 1 illustrates these goals with two backpacks. Finding the blue backpack's own description requires distinguishing it from the red one, whereas searching for other backpacks makes their shared category relevant. Co-purchases can connect both similar and complementary products. Pulling related objects together may improve cross-object retrieval but make a true image-text pair harder to distinguish from similar alternatives. The challenge is to use relational agreement without losing the object-specific information required for paired matching.

Paired ranking and contrastive objectives learn image-text correspondence, while generalized retrieval objectives extend supervision to richer definitions of relevance (Faghri et al., 2017; Radford et al., 2021; Zhai et al., 2023; Yamaguchi et al., 2025; Zhu et al., 2024; Lee et al., 2025). Multimodal graph models add neighborhood aggregation, filtering, and modality-topology interaction (Ning et al., 2025; Guo et al., 2024; Zheng et al., 2025; Su et al., 2026; Zhang et al., 2026; Zuo et al., 2026). These methods support correspondence or relational information sharing, but they do not determine where graph supervision should act when one set of learned weights must serve both retrieval goals.

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.

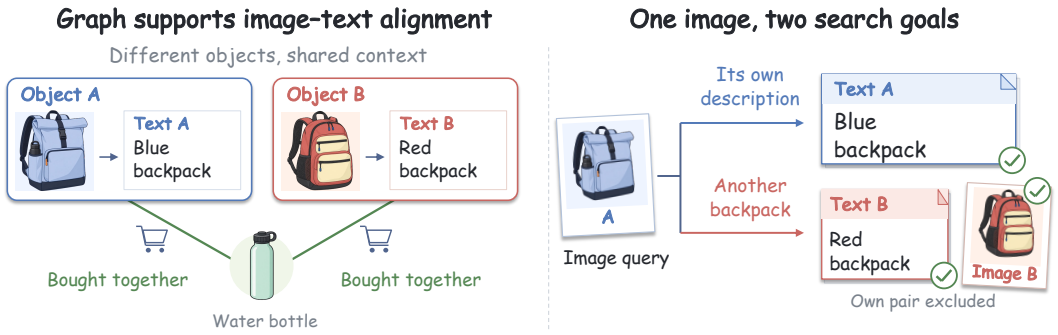

Figure 1: Different retrieval goals within one collection. Identity matches an object's own attributes, semantic retrieval seeks other objects in the same category, and observed relations connect distinct objects.

Graph signal processing provides a precise view of this tension. Graph smoothness measures how representations vary across observed relationships (Shuman et al., 2013). Relational agreement favors low variation, while paired identity still requires differences between objects. Intermediate and projected representations can also retain different information (Chen et al., 2020; Wu et al., 2026). The question is which representation graph smoothness should regulate and how its learned signal should guide later propagation.

Our empirical study verifies category information beyond degree-preserving graph relabelings, identifies an identity-semantic tradeoff in a shared output, and shows that uniform graph propagation reduces part of the gain supplied by graph regularization. These observations suggest two connected design principles. Identity and relational semantics should remain accessible through separately supervised embeddings, and the learned semantic signal should determine how information is shared during graph propagation.

To this end, we propose GOMA. Each modality produces a content embedding for paired identity and a semantic embedding learned from cross-modal pairs and observed relationships. Single-modality retrieval uses the embedding that matches its target. For complete records, both embeddings form the initial fused representation $\mathbf{H}_0$, semantic agreement weights observed edges, and restart graph propagation reinjects $\mathbf{H}_0$ at every step. The semantic embedding therefore connects graph regularization to propagation by controlling the exchange between neighbors.

Our Contributions. (1) *Valuable Insights.* We show that retrieval relevance determines where graph smoothness should act and why graph regularization and uniform graph propagation need not provide additive gains. (2) *Novel Method.* We propose GOMA, which jointly learns separately supervised content and semantic embeddings, uses semantic agreement to weight observed edges, and applies restart graph propagation to complete records. One set of learned weights supports four retrieval tasks through a task-specific readout. (3) *SOTA Performance.* Across six datasets and four retrieval tasks, GOMA leads all 14 primary measures against 14 external methods. Core removals, equal-capacity comparisons, and signal measurements identify the roles of its components.

2 Problem Formulation

We first define the data and observed graph, followed by the retrieval interfaces and relevance rules. This separates model inputs from evaluation targets.

2.1 Data and Graph Setup

Let $\mathcal{V}$ contain $N = |\mathcal{V}|$ objects and let $\mathcal{M} = \{v, t\}$ denote the image and text modalities. Object $i \in \mathcal{V}$ has paired attributes represented by prepared features $\mathbf{x}_i^m \in \mathbb{R}^d$ for each $m \in \mathcal{M}$. An observed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ records relationships between distinct objects, such as co-purchases or interactions. The paired attributes and graph are model inputs. Category labels and held-out relationships define analysis or evaluation targets.

2.2 Retrieval Interfaces

The inference input determines which information may be used for scoring. In *single-modality retrieval*, an image query is compared with text candidates, or a text query with image candidates. Every object is encoded from its own supplied attribute, without its missing attribute or an evaluation-time neighborhood. In *dual-attribute retrieval*, both attributes and the permitted observed graph are available for graph propagation and offline index construction. The interface defines the available information, while the rules below define a correct candidate.

2.3 Retrieval Tasks

For a query object i, the system ranks candidate objects j from the prescribed candidate set. The four tasks combine the two interfaces with three definitions of relevance.

Single-modality identity and semantic retrieval. Identity retrieval asks for the opposite-modality attribute of the same object. Candidate j is relevant to query i exactly when $j = i$. We report recall at ranks 1, 5, and 10 in both directions, with bidirectional R@1 as the primary measure. Cross-object semantic retrieval instead seeks another object in the same category. Let y_i denote the evaluation category of object i. A candidate is relevant when $y_j = y_i$ and $j \neq i$. Query and gallery objects are disjoint, so the query's paired attribute cannot satisfy this task. The primary measure is category-macro mAP@10, which first averages query scores within each category and then gives all categories equal weight.

Dual-attribute semantic retrieval and relation ranking. Semantic retrieval keeps the same cross-object relevance rule while allowing both attributes and the permitted observed graph. Its primary measure remains category-macro mAP@10. Relation ranking asks the model to recover a missing relationship by ranking the true linked object among prescribed negative candidates. Its primary measure is mean reciprocal rank (MRR). Before propagation, each target relationship and its reverse orientation are removed from the supplied graph. Its endpoints may retain other observed neighbors, so the task measures whether the remaining context helps recover the missing link.

2.4 Training and Evaluation Setting

Let $\mathcal{T} \subset \mathcal{V}$ be the training objects and let $\mathcal{E}_{\mathcal{T}}$ contain the observed edges whose endpoints both lie in $\mathcal{T}$. Learning uses their paired attributes and the induced graph $(\mathcal{T}, \mathcal{E}_{\mathcal{T}})$. Category labels are used only for analysis and evaluation and never supervise GOMA. Hidden relation targets are removed from the observed graph before propagation. The goal is one set of learned parameters per dataset and seed that supports every applicable task. Sec. 5 specifies model selection, and Appendix E gives the partitions, candidate pools, and metric averaging.

3 Empirical Motivation

We examine relational information, retrieval targets, and the interaction between graph regularization and propagation. Graph analysis uses training objects and categories, representation analysis uses validation candidates, and the propagation probe uses matched training steps on the established evaluation pools. Appendix B.4 gives complete definitions and measurements.

Relationships contain semantic information. Fig. 2(a) compares same-category training-edge proportions with three vertex relabelings within degree groups. Observed agreement is higher on all four category datasets by 22.67 to 62.97 percentage points. Because the control retains each object's training degree, the difference shows category structure beyond that explained by degree and category frequencies. The graph can therefore supply cross-object information that is useful to investigate for semantic retrieval.

Different targets favor different embeddings. The shared-output probe has two residual mappings per modality, with identity, paired, and graph objectives at the final output. At the same weights, the final output improves semantic mAP@10 on all four datasets but reduces identity R@1 on three (Fig. 2(b)). Books improves in both. The Books exception indicates that the shared-output trade-off is dataset-dependent rather than universal. Graph regularization still updates the content mapping through the semantic branch, so this comparison concerns supervision placement. We therefore

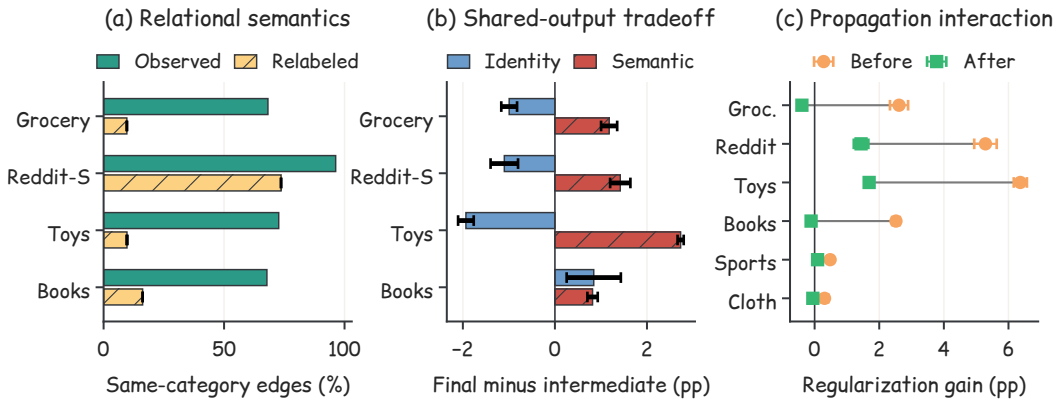

Figure 2: Graph semantics, shared-output retrieval, and propagation. Error bars give graph-relabeling SD in (a) and paired model-seed SD in (b,c).

keep identity supervision at the content embedding and use the semantic embedding for cross-object retrieval.

Graph regularization and propagation can interact negatively. The last probe crosses graph regularization and uniform graph propagation at the same training step and with the same initialization rule. Graph regularization improves the pre-propagation retrieval score on all six datasets, but uniform propagation reduces part of that gain in every seed (Fig. 2(c)). Because propagation still helps both training variants, the result identifies overlapping gains rather than universal harm. It motivates using the learned semantics to guide propagation.

4 Methodology

Fig. 3 organizes GOMA into joint representation learning and semantic-guided graph propagation. Each modality produces content and semantic embeddings. The learned semantics guide propagation for complete records, connecting one model to all four retrieval tasks.

4.1 Content and Semantic Embeddings

Let $\mathcal{N}$ normalize each row to unit length and let $m \in \{v, t\}$ denote the modality. From normalized input $\mathbf{x}_i^m$, GOMA computes

$$\mathbf{c}_i^m = \mathcal{N}(\mathbf{x}_i^m + f_m(\mathbf{x}_i^m)), \qquad \mathbf{z}_i^m = \mathcal{N}(\mathbf{c}_i^m + g_m(\mathbf{c}_i^m)). \tag{1}$$

The two-layer mappings f_m and g_m have width d and a GELU activation. Their final linear layers start at zero, initially retaining the supplied content. The content embedding $\mathbf{c}_i^m$ serves object identity, while $\mathbf{z}_i^m$ expresses relationships between objects. Each computation uses only its own modality.

The content embedding receives bidirectional VSE++ hard-negative ranking loss $\mathcal{L}_{\text{id}}$ with margin 0.2, separating each true pair from its most competitive incorrect candidate (Faghri et al., 2017). At the semantic embedding, bidirectional paired cross-entropy $\mathcal{L}_{\text{pair}}$ with temperature 0.1 maintains image-text correspondence. Both objectives update the content mapping. Identity remains directly supervised while the jointly learned semantic transformation expresses cross-object agreement.

4.2 Graph-Regularized Training

Graph smoothness encourages the semantic outputs to agree across training relationships. Let d_i be the degree of training object i in the complete induced training graph. We minimize the degree-normalized edge energy

$$\mathcal{E}_{\mathcal{G}}(\mathbf{Z}^v, \mathbf{Z}^t) = \frac{1}{2d|\mathcal{E}_{\mathcal{T}}|} \sum_{m \in \{v,t\}} \sum_{(i,j) \in \mathcal{E}_{\mathcal{T}}} \left\| \frac{\mathbf{z}_i^m}{\sqrt{d_i}} - \frac{\mathbf{z}_j^m}{\sqrt{d_j}} \right\|_2^2. \tag{2}$$

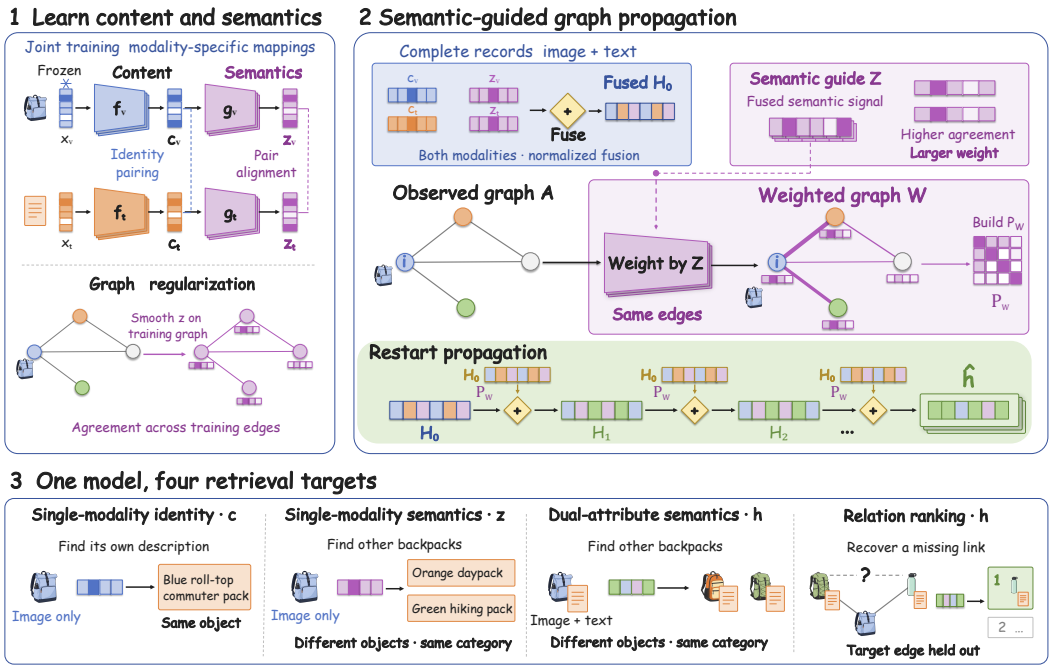

Figure 3: GOMA jointly learns content and semantic embeddings. For complete records, both form the initial fused representation and the learned semantics guide graph propagation. One model supports all four retrieval tasks.

Here $\mathbf{Z}^m$ stacks the semantic embeddings and the edge list contains both orientations. Degree normalization accounts for the different numbers of relationships attached to objects. Isolated training objects contribute through paired supervision. This is the normalized graph Dirichlet principle applied within each modality (Shuman et al., 2013). Cross-modal pairing establishes correspondence, and graph energy transfers the observed relational structure into independently computable attribute representations.

The combined objective is

$$\mathcal{L} = \mathcal{L}_{\text{id}} + \eta \left(\mathcal{L}_{\text{pair}} + \lambda \mathcal{E}_{\mathcal{G}} \right). \tag{3}$$

We set η and λ once from gradient norms on ten training batches at initialization, then retain them throughout learning. Each update samples paired objects and training edges independently. Appendix D.1 gives the exact objectives and loss-weight construction. The observed graph influences single-modality retrieval through parameters learned from the training graph. At inference, each query encoder uses only its supplied attribute.

4.3 Semantic-Guided Graph Propagation

For complete records, we fuse the two attributes into content $\mathbf{C} = \mathcal{N}(\mathbf{C}^v + \mathbf{C}^t)$ and semantics $\mathbf{Z} = \mathcal{N}(\mathbf{Z}^v + \mathbf{Z}^t)$, where each matrix stacks the corresponding embedding. Their equal fusion $\mathbf{H}_0 = \mathcal{N}(\mathbf{C} + \mathbf{Z})$ defines the initial signal for graph propagation.

The semantic signal also describes which neighbors agree in the learned retrieval space. For each allowed observed edge, we set

$$W_{ij} = A_{ij} \frac{\sigma^2}{\sigma^2 + \|\mathbf{Z}_i - \mathbf{Z}_j\|_2^2}. \tag{4}$$

Here $\mathbf{A}$ is the binary symmetric adjacency of the permitted observed graph. The scale σ^2 is the mean squared semantic distance over legal training edges, with a numerical floor of 10^{-8}. Close semantic neighbors receive larger weights, while every observed edge retains positive weight. Similarity-based weighting is an established way to guide graph aggregation (Yang et al., 2021). GOMA uses it to

connect the learned semantic representation to graph propagation. The weights are computed once and remain unchanged during propagation.

In the example of Fig. 1, the graph can connect similar backpacks and complementary accessories. Learned semantic distances adjust their influence while retaining both relationships. Graph regularization thus affects propagation through both the initial signal and the edge weights.

Let $\mathbf{D}_W$ contain weighted degrees and $\mathbf{P}_W = \frac{1}{2}(\mathbf{I} + \mathbf{D}_W^{-1/2}\mathbf{W}\mathbf{D}_W^{-1/2})$. Isolated objects use $(P_W)_{ii} = 1$. GOMA propagates the initial signal through

$$\mathbf{H}_{k+1} = \alpha\mathbf{H}_0 + (1-\alpha)\mathbf{P}_W\mathbf{H}_k, \qquad k = 0, \ldots, K-1. \tag{5}$$

We use $\alpha = 0.1$ and $K = 8$ for every dataset, followed by row normalization $\widehat{\mathbf{H}} = \mathcal{N}(\mathbf{H}_K)$. The restart term $\alpha\mathbf{H}_0$ reinjects the initial fused representation at every step (Klicpera et al., 2019). For the computed operator, the recurrence is a finite low-pass graph filter that retains a positive contribution from the initial signal at every frequency. Final row normalization is a separate nonlinear operation. Appendix C derives the filter and distinguishes these two steps.

4.4 Retrieval Scoring

The required relevance determines the score. For image-to-text and dual-attribute retrieval, GOMA uses

$$s(i,j) = \begin{cases} (\mathbf{c}_i^v)^\top \mathbf{c}_j^t, & \text{single-modality identity,} \\ (\mathbf{z}_i^v)^\top \mathbf{z}_j^t, & \text{single-modality semantic,} \\ \widehat{\mathbf{h}}_i^\top \widehat{\mathbf{h}}_j, & \text{dual-attribute semantic or relation.} \end{cases} \tag{6}$$

Text-to-image retrieval exchanges v and t. The first two scores encode attributes independently, while the third uses complete records and permitted relationships. All tasks share learned weights. Offline index construction costs $O(|\mathcal{E}|d)$ for edge weighting and $O(K(|\mathcal{E}| + N)d)$ for propagation. Subsequent ranking uses cached representations.

The retrieval interface determines the choice among $\mathbf{c}$, $\mathbf{z}$, and $\widehat{\mathbf{h}}$. This task-specific readout distinguishes identity, cross-object semantic, and relation relevance.

5 Experiments

5.1 Experimental Setup

Datasets and tasks. Grocery, Reddit-S, Toys, and Books support identity and cross-object semantic retrieval. Sports and Cloth support relation ranking. These multimodal graph collections contain 15,894 to 685,294 objects (Yan et al., 2024; Wan et al., 2026). All methods receive the same prepared features, object partitions, candidates, and relevance rules. The two semantic interfaces share query/gallery partitions and differ in their available attributes and graph information.

External methods. We compare the same 14 methods in both interfaces. Vision-language methods are VSE++, CLIP-Refine, and SigLIP (Faghri et al., 2017; Yamaguchi et al., 2025; Zhai et al., 2023). Graph encoders are GCN, GAT, and GraphSAGE (Kipf & Welling, 2017; Veličković et al., 2018; Hamilton et al., 2017). Multimodal graph methods are MMGCN, MGAT, LGMRec, DMGC, DGF, CAMPA, CoMAG, and RoleMAG (Wei et al., 2019; Tao et al., 2020; Guo et al., 2024; 2025; Zheng et al., 2025; Su et al., 2026; Zhang et al., 2026; Zuo et al., 2026). Their adaptations retain their core computations and losses on the common feature inputs. Graph methods use their graph or fusion outputs for dual-attribute retrieval and independently computable missing-input views for single-modality queries. Appendix F describes these interfaces. VSE++, CLIP-Refine, and SigLIP are direct vision-language comparators for single-modality identity retrieval. GCN, GAT, GraphSAGE, and the multimodal graph methods are missing-modality adaptation controls that retain their graph structure and auxiliary objectives while using independently computable views.

Selection and reporting. We report percentage means and sample SD over seeds 43, 44, and 45. GOMA uses AdamW with learning rate 0.001, weight decay 0.0001, and at most 500 updates, validating every five updates with 50-update patience. Category datasets select by the mean of identity R@1 and both semantic mAP@10 measures, while relation datasets select by MRR. One

GOMA model per dataset and seed serves all tasks. External methods retain their task-specific selectors. Bold and underline mark the best and second-best displayed means. Appendix B includes all metrics and directional recalls.

5.2 Performance Comparison

GOMA leads both primary single-modality measures on all four datasets in Table 1. Semantic mAP@10 gains over the strongest external method range from 3.33 to 8.85 points, while identity R@1 improves by 0.17 to 0.91 points. The same model supplies both outputs. These results show that training relationships can improve cross-object search while retaining a strong representation for paired identity retrieval.

Table 1: Single-modality identity R@1 and semantic mAP@10. Each query and candidate uses only its own supplied attribute.

Method	Grocery		Reddit-S		Toys		Books	
	Id. R@1	Sem. mAP	Id. R@1	Sem. mAP	Id. R@1	Sem. mAP	Id. R@1	Sem. mAP
Vision-language learning								
VSE++	49.94 ± 0.10	30.53 ± 0.14	16.10 ± 0.40	37.34 ± 0.26	59.19 ± 0.12	23.09 ± 0.09	<u>39.23</u> ± 0.09	39.99 ± 0.29
CLIP-Refine	49.93 ± 0.07	33.72 ± 0.06	<u>18.48</u> ± 0.11	46.19 ± 0.10	<u>62.57</u> ± 0.10	27.57 ± 0.05	35.44 ± 0.01	<u>42.82</u> ± 0.02
SigLIP	<u>52.36</u> ± 0.30	<u>38.27</u> ± 0.12	17.14 ± 0.19	<u>50.89</u> ± 0.04	61.29 ± 0.49	<u>32.75</u> ± 0.25	37.84 ± 0.07	42.42 ± 0.07
Graph encoders								
GCN	4.60 ± 0.13	8.86 ± 0.53	0.97 ± 0.24	11.16 ± 1.20	2.87 ± 0.31	4.55 ± 0.06	0.32 ± 0.04	16.65 ± 0.99
GAT	5.97 ± 1.07	11.96 ± 0.42	1.20 ± 0.17	16.41 ± 1.46	4.64 ± 0.78	5.08 ± 0.49	0.91 ± 0.19	23.08 ± 0.67
GraphSAGE	1.82 ± 0.38	6.78 ± 0.45	0.56 ± 0.16	10.17 ± 1.26	1.17 ± 0.21	3.46 ± 0.17	0.38 ± 0.07	17.29 ± 1.20
Multimodal graph learning								
MMGCN	15.33 ± 1.13	28.15 ± 0.53	4.42 ± 0.24	42.22 ± 0.36	14.26 ± 0.97	16.94 ± 0.76	12.44 ± 0.32	40.82 ± 0.73
MGAT	20.82 ± 0.29	30.54 ± 1.00	5.92 ± 0.28	43.60 ± 0.98	22.81 ± 0.73	20.51 ± 0.31	17.23 ± 0.54	41.71 ± 0.12
LGMRec	9.39 ± 0.60	12.88 ± 0.23	2.16 ± 0.41	18.79 ± 0.39	8.61 ± 0.11	6.50 ± 0.43	1.54 ± 0.02	25.63 ± 0.98
DMGC	24.05 ± 0.95	34.70 ± 0.23	6.88 ± 0.31	50.60 ± 0.27	23.74 ± 0.68	17.48 ± 0.35	10.36 ± 0.47	42.14 ± 0.37
DGF	20.29 ± 0.62	32.45 ± 0.96	5.71 ± 0.48	44.62 ± 0.97	20.81 ± 0.41	23.01 ± 0.64	13.11 ± 0.52	40.89 ± 0.27
CAMPA	3.25 ± 1.15	8.76 ± 0.78	0.73 ± 0.14	10.52 ± 1.77	3.69 ± 0.75	5.37 ± 0.52	0.24 ± 0.06	16.07 ± 1.49
CoMAG	12.15 ± 0.87	13.84 ± 0.75	2.34 ± 0.34	17.05 ± 1.51	11.20 ± 0.91	7.18 ± 0.22	1.46 ± 0.16	26.37 ± 0.61
RoleMAG	5.57 ± 0.25	9.32 ± 0.31	0.98 ± 0.11	13.06 ± 0.65	4.07 ± 0.89	4.93 ± 0.33	0.61 ± 0.06	20.76 ± 0.85
Proposed method								
GOMA	**52.96** ± 0.06	**42.82** ± 0.22	**18.66** ± 0.19	**59.74** ± 0.26	**62.94** ± 0.40	**41.52** ± 0.44	**40.14** ± 0.31	**46.16** ± 0.03

Table 2 gives the six dual-attribute primary results. GOMA improves semantic mAP@10 by 3.71 to 12.86 points and relation MRR by 0.39 to 0.78 points over the strongest external methods. It also leads all 18 dual-attribute metrics and all 12 single-modality semantic metrics in the complete tables. Together, the two interfaces yield state-of-the-art results on all 14 primary measures and 41 of the 42 aggregate measures. The component studies below examine how the learned representations and observed graph produce these results.

5.3 Ablation Study

The w/o graph regularization variant removes $\mathcal{E}_{\mathcal{G}}$ while retaining both embeddings, paired objectives, and semantic-guided propagation. The w/o graph propagation variant scores $\mathbf{H}_0$ directly. Both are retrained under the same validation rule.

Both removals reduce the primary metric on five datasets. On Cloth, graph propagation remains essential, while graph regularization produces a small MRR decrease.

Table 4 separates supervision placement from capacity. Shared readout keeps all parameters but applies identity supervision and scoring at $\mathbf{z}$. GOMA gains 0.53 to 2.04 points in identity R@1 and 3.05 to 7.49 in semantic mAP@10. Removing the semantic mapping ($\mathbf{z} = \mathbf{c}$) also lowers both primary measures on every category dataset, supporting the two separately supervised embeddings.

Table 2: Dual-attribute retrieval across six datasets. The first four report semantic mAP@10 and the last two report relation MRR.

Method	Cross-object semantic retrieval				Relation ranking	
	Grocery mAP@10	**Reddit-S** mAP@10	**Toys** mAP@10	**Books** mAP@10	**Sports** MRR	**Cloth** MRR
Vision-language learning						
VSE++	39.67 ± 0.13	56.19 ± 0.39	29.58 ± 0.64	46.23 ± 0.43	62.61 ± 0.77	59.20 ± 0.87
CLIP-Refine	34.81 ± 0.32	59.37 ± 2.67	25.79 ± 0.76	49.98 ± 0.02	67.15 ± 0.02	62.29 ± 0.02
SigLIP	44.17 ± 0.08	62.95 ± 0.11	38.44 ± 0.10	48.84 ± 0.13	67.09 ± 0.16	61.36 ± 1.64
Graph encoders						
GCN	36.50 ± 1.06	64.61 ± 0.86	36.15 ± 0.93	62.90 ± 0.21	87.50 ± 0.18	80.34 ± 0.11
GAT	35.67 ± 0.64	67.93 ± 0.67	37.76 ± 0.99	62.03 ± 0.23	86.67 ± 0.07	79.48 ± 0.12
GraphSAGE	28.95 ± 1.89	58.01 ± 1.22	25.90 ± 1.64	61.06 ± 0.40	82.05 ± 0.42	75.33 ± 0.03
Multimodal graph learning						
MMGCN	35.21 ± 0.82	64.74 ± 1.29	26.84 ± 0.65	61.14 ± 0.57	75.48 ± 0.43	67.29 ± 0.29
MGAT	40.65 ± 0.75	67.43 ± 1.69	29.65 ± 0.04	60.58 ± 0.43	73.58 ± 0.41	66.23 ± 0.45
LGMRec	41.80 ± 1.64	69.29 ± 0.44	36.89 ± 0.35	62.99 ± 0.40	87.07 ± 0.11	80.54 ± 0.32
DMGC	32.56 ± 11.45	55.13 ± 1.47	13.53 ± 0.58	48.56 ± 0.23	57.38 ± 16.31	54.10 ± 13.77
DGF	34.57 ± 0.54	61.93 ± 0.14	27.34 ± 3.98	59.93 ± 0.32	77.06 ± 0.37	71.91 ± 0.85
CAMPA	30.12 ± 1.15	56.32 ± 0.59	24.53 ± 1.37	47.21 ± 1.07	64.07 ± 0.34	56.14 ± 2.12
CoMAG	40.31 ± 1.91	67.28 ± 1.53	34.36 ± 2.69	59.38 ± 1.28	84.54 ± 0.64	74.33 ± 2.55
RoleMAG	38.11 ± 0.32	65.43 ± 1.21	36.35 ± 1.02	62.26 ± 0.28	85.20 ± 0.83	76.83 ± 0.52
Proposed method						
GOMA	**49.60** ± 0.05	**78.97** ± 0.05	**51.31** ± 0.07	**66.70** ± 0.01	**87.89** ± 0.03	**81.32** ± 0.11

Table 3: Core graph ablations across all six dual-attribute datasets.

Variant	**Grocery**	**Reddit-S**	**Toys**	**Books**	**Sports**	**Cloth**
GOMA	**49.60** ± 0.05	**78.97** ± 0.05	**51.31** ± 0.07	**66.70** ± 0.01	**87.89** ± 0.03	81.32 ± 0.11
w/o graph regularization	49.55 ± 0.09	77.24 ± 0.10	49.18 ± 0.07	66.62 ± 0.03	87.78 ± 0.02	**81.46** ± 0.00
w/o graph propagation	47.46 ± 0.25	69.81 ± 0.18	45.65 ± 0.41	51.01 ± 0.14	68.24 ± 0.11	63.36 ± 0.15

Table 4: Separate supervision and semantic mapping for single-modality retrieval.

Variant	**Grocery**		**Reddit-S**		**Toys**		**Books**	
	Id. R@1	Sem. mAP	Id. R@1	Sem. mAP	Id. R@1	Sem. mAP	Id. R@1	Sem. mAP
GOMA	**52.96** ± 0.06	**42.82** ± 0.22	**18.66** ± 0.19	**59.74** ± 0.26	**62.94** ± 0.40	**41.52** ± 0.44	**40.14** ± 0.31	**46.16** ± 0.03
Shared readout	51.79 ± 0.13	39.68 ± 0.16	18.13 ± 0.30	52.26 ± 0.45	60.89 ± 0.23	37.36 ± 0.60	39.00 ± 0.10	43.11 ± 0.10
w/o semantic mapping	51.83 ± 0.12	39.53 ± 0.29	18.21 ± 0.31	52.71 ± 0.47	60.86 ± 0.17	37.43 ± 0.14	39.04 ± 0.07	43.05 ± 0.14

Graph regularization improves single-modality semantic mAP@10 by 2.13 to 6.80 points. Relabeling its edges lowers this measure on every category dataset, showing that the connected objects matter beyond paired supervision. Appendix B.2 gives all representation, training, and initialization controls.

5.4 Representation Analysis

Fig. 4 shows that the semantic embedding reduces training-edge variation and improves semantic mAP@10 by 6.0–19.1 points, while the content embedding ranks more true pairs. The equal-capacity controls support these target-specific roles.

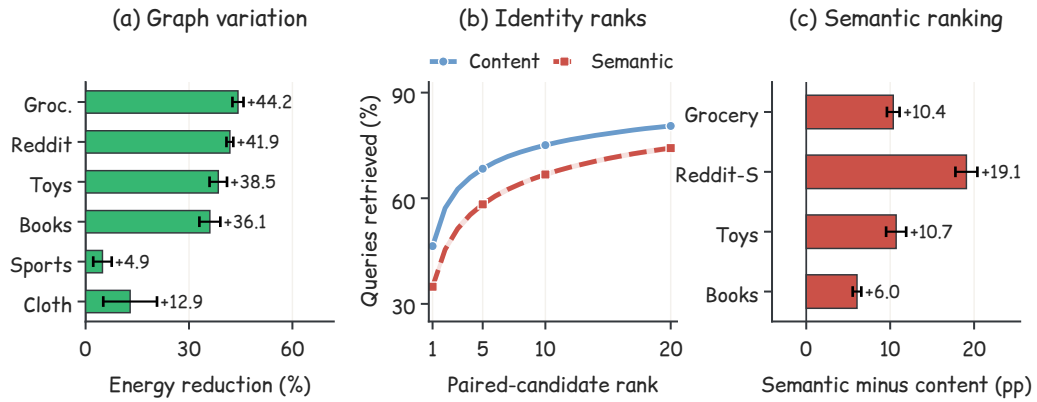

Figure 4: GOMA's embeddings at the same weights. Training-edge variation, validation identity ranks, and semantic mAP@10 distinguish their roles. Bars and bands show sample SD.

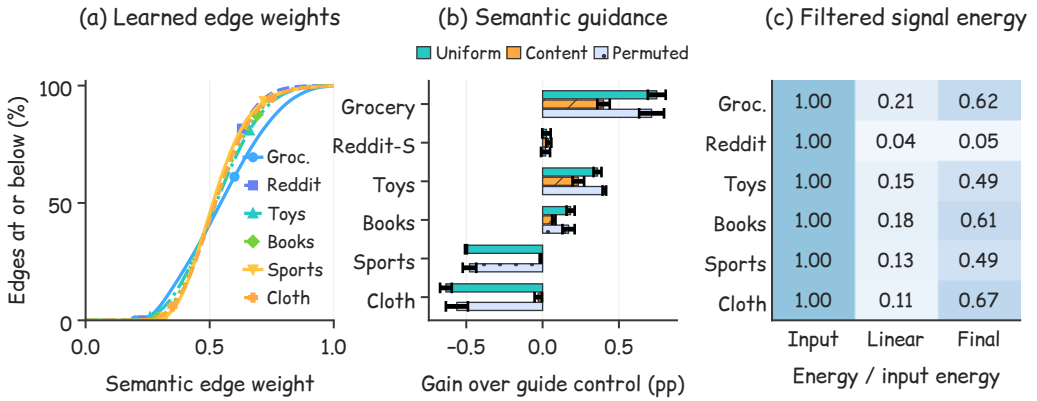

Figure 5: Semantic-guided graph propagation. Edge-weight distributions, primary-metric guide comparisons, and relative graph energy cover all six datasets.

5.5 Graph Propagation Analysis

Fig. 5(a) shows heterogeneous edge weights, and panel (b) changes only the guide at the same checkpoint. GOMA improves over uniform, content-based, and permuted-semantic guides on the four category datasets, while Sports and Cloth favor simpler guides. Panel (c) shows that propagation lowers graph energy while retaining alignment with $\mathbf{H}_0$. Appendix B.3 gives complete results.

6 Conclusion

Multimodal retrieval over graph-structured collections must preserve object identity while using relationships for cross-object search. We showed that these goals favor different representations and that uniform graph propagation can weaken graph-regularization gains. GOMA therefore learns separate content and semantic embeddings, uses semantics to weight observed edges, and reinjects the initial fused representation during propagation. A task-specific readout selects the representation that matches each relevance definition, and the propagated representations form an offline retrieval index. Across six datasets and four tasks, GOMA leads all 14 primary measures against 14 external methods. Ablations and signal measurements explain the roles of supervision placement, graph regularization, and graph propagation, and show that simpler edge weights can suit relation ranking. Overall, graph information should be organized around the retrieval target rather than forced through one shared output.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of this manuscript, including grammar, phrasing, and clarity, as the authors are non-native English speakers. We have reviewed all AI-assisted work, and the authors verified the final text against the reported results and claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The problem formulation, method, and appendix specify the graph operators, training objective, sampling and model selection, data construction, feature resources, evaluation protocol, and the adaptations applied to external methods. The mathematical analysis states the assumptions behind the reported claims, and the appendix reports the complete retrieval results, ablations, and computational cost. Anonymous code will accompany the submission.

REFERENCES

Song Bai, Xiang Bai, Qi Tian, and Longin Jan Latecki. Regularized Diffusion Process for Visual Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. URL https://ojs.aaai.org/index.php/AAAI/article/view/11216.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Ziang Chen, Zhengjiang Lin, Shi Chen, Yury Polyanskiy, and Philippe Rigollet. Residual connections provably mitigate oversmoothing in graph neural networks. *arXiv preprint arXiv:2501.00762*, 2025. URL https://arxiv.org/abs/2501.00762.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

Dongzhe Fan, Yi Fang, Jiajin Liu, Djellel Difallah, and Qiaoyu Tan. Mlaga: Multimodal large language and graph assistant. *CoRR*, abs/2506.02568, 2025.

Yi Fang, Bowen Jin, Jiacheng Shen, Sirui Ding, Qiaoyu Tan, and Jiawei Han. GRAPHGPT-O: synergistic multimodal comprehension and generation on graphs. In *CVPR*, pp. 19467–19476. Computer Vision Foundation / IEEE, 2025.

Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. Learning semantic relationship among instances for image-text matching. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 15159–15168, 2023.

Zheren Fu, Lei Zhang, Hou Xia, and Zhendong Mao. Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 26307–26316, June 2024.

Zhaochen Guo, Zhixiang Shen, Xuanting Xie, Liangjian Wen, and Zhao Kang. Disentangling homophily and heterophily in multimodal graph clustering. *CoRR*, abs/2507.15253, 2025.

Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. Lgmrec: Local and global graph learning for multimodal recommendation. In *AAAI*, pp. 8454–8462. AAAI Press, 2024.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, pp. 1024–1034, 2017.

Yufei He, Yuan Sui, Xiaoxin He, Yue Liu, Yifei Sun, and Bryan Hooi. Unigraph2: Learning a unified embedding space to bind multimodal graphs. In *WWW*, pp. 1759–1770. ACM, 2025.

Jun Hu, Bryan Hooi, Bingsheng He, and Yinwei Wei. Modality-independent graph neural networks with global transformers for multimodal recommendation. In *AAAI*, pp. 11790–11798. AAAI Press, 2025.

Jun Hu, Yufei He, Yuan Li, Bryan Hooi, and Bingsheng He. NTSFormer: A self-teaching graph transformer for multimodal isolated cold-start node classification. *CoRR*, abs/2507.04870, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/38506.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondrej Chum. Efficient Diffusion on Region Manifolds: Recovering Small Objects With Compact CNN Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Iscen_Efficient_Diffusion_on_CVPR_2017_paper.html.

Adarsh Jamadandi, Celia Rubio-Madrigal, and Rebekka Burkholz. Spectral Graph Pruning Against Over-Squashing and Over-Smoothing. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2404.04612.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/04261fce1705c4f02f062866717d592a-Abstract-Conference.html.

Vassilis Kalofolias, Xavier Bresson, Michael Bronstein, and Pierre Vandergheynst. Matrix completion on graphs. *arXiv preprint arXiv:1408.1717*, 2014. URL https://arxiv.org/abs/1408.1717.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR (Poster)*. OpenReview.net, 2017.

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gL-2A9Ym.

Jungsoo Lee, Janghoon Cho, Hyojin Park, Munawar Hayat, Kyuwoong Hwang, Fatih Porikli, and Sungha Choi. Generalized Contrastive Learning for Universal Multimodal Retrieval. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2509.25638.

Lei Li, Souhaib Attaiki, and Maks Ovsjanikov. SRFeat: Learning locally accurate and globally consistent non-rigid shape correspondence. In *International Conference on 3D Vision (3DV)*, 2022. URL https://arxiv.org/abs/2209.07806.

Xunkai Li, Zekai Chen, Zhengyu Wu, Henan Sun, Daohan Su, Guang Zeng, Hongchao Qin, Rong-Hua Li, and Guoren Wang. LION: A Clifford Neural Paradigm for Multimodal-Attributed Graph Learning. *arXiv preprint arXiv:2608.24795*, 2026. URL https://arxiv.org/abs/2608.24795.

Zihan Liang, Yufei Ma, ZhiPeng Qian, Huangyu Dai, Zihan Wang, Ben Chen, Chenyi Lei, Yuqing Ding, and Han Li. UniECS: Unified Multimodal E-Commerce Search Framework with Gated Cross-modal Fusion. *arXiv preprint arXiv:2508.13843*, 2025. URL https://arxiv.org/abs/2508.13843.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-Embed: Universal Multimodal Retrieval with Multimodal LLMs. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2411.02571.

Zengrong Lin, Zheng Wang, Tianwen Qian, Pan Mu, Sixian Chan, and Cong Bai. NeighborRetr: Balancing Hub Centrality in Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b. URL https://arxiv.org/abs/2503.10526.

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *CVPR*, pp. 10918–10927. Computer Vision Foundation / IEEE, 2020.

Jiajin Liu, Dongzhe Fan, Jiacheng Shen, Chuanhao Ji, Daochen Zha, and Qiaoyu Tan. Graph-mllm: Harnessing multimodal large language models for multimodal graph learning. *CoRR*, abs/2506.10282, 2025a.

Yang Liu, Wentao Feng, Zhuoyao Liu, Shudong Huang, and Jiancheng Lv. Aligning information capacity between vision and language via dense-to-sparse feature distillation for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21679–21688, October 2025b.

Yang Liu, Mengyuan Liu, shudong Huang, and Jiancheng Lv. Asymmetric visual semantic embedding framework for efficient vision-language alignment. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence*, 2025c.

Yang Liu, Wentao Feng, Shu-Dong Huang, Yalan Ye, and Jiancheng Lv. Intra-modal neighbors never lie: Rectifying inter-modal noisy correspondence via graph-based intra-modal reasoning. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.04061.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net, 2019.

Xuying Ning, Dongqi Fu, Tianxin Wei, Wujiang Xu, and Jingrui He. Graph4mm: Weaving multimodal learning with structural information. In *ICML*. OpenReview.net, 2025.

MoonJeong Park and Dongwoo Kim. Taming Gradient Oversmoothing and Expansion in Graph Neural Networks. *arXiv preprint arXiv:2410.04824*, 2024. URL https://arxiv.org/abs/2410.04824.

Moonjeong Park, Jaeseung Heo, and Dongwoo Kim. Mitigating Oversmoothing Through Reverse Process of GNNs for Heterophilic Graphs. In *Proceedings of the International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/park24d.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael T. Schaub. Residual Connections and Normalization Can Provably Prevent Oversmoothing in GNNs. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6c473e69ba261200dd595d07494c1a73-Abstract-Conference.html.

David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 2013. doi: 10.1109/MSP.2012.2235192.

Daohan Su, Hao Liu, Xunkai Li, Yinlin Zhu, Yongfu Xiong, Yi Liu, Hongchao Qin, Rong-Hua Li, and Guoren Wang. CAMPA: Efficient and aligned multimodal graph learning via decoupled propagation and aggregation. *arXiv preprint arXiv:2605.11468*, 2026.

Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. MGAT: multimodal graph attention network for recommendation. *Inf. Process. Manag.*, 57(5):102277, 2020.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Chenxi Wan, Xunkai Li, Yilong Zuo, Haokun Deng, Sihan Li, Bowen Fan, Hongchao Qin, Rong-Hua Li, and Guoren Wang. OpenMAG: A comprehensive benchmark for multimodal-attributed graph. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64650.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and Benchmarking Universal Multimodal Information Retrievers. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2311.17136.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video. In *ACM Multimedia*, pp. 1437–1445. ACM, 2019.

Zhengyu Wu, Xu Wang, Hongchao Qin, Xunkai Li, Guang Zeng, Rong-Hua Li, and Guoren Wang. SMGFM: Spectral multimodal graph pretraining for multimodal-attributed graphs. *arXiv preprint arXiv:2606.12867*, 2026.

Shaoan Xie, Lingjing Kong, Yujia Zheng, Yu Yao, Zeyu Tang, Eric P. Xing, Guangyi Chen, and Kun Zhang. Smartclip: Modular vision-language alignment with identification guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29780–29790, June 2025.

Haobo Xu, Yuchen Yan, Dingsu Wang, Zhe Xu, Zhichen Zeng, Tarek F. Abdelzaher, Jiawei Han, and Hanghang Tong. SLOG: An Inductive Spectral Graph Neural Network Beyond Polynomial Filter. In *Proceedings of the International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/xu24aa.html.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith C. H. Ngai. MENTOR: Multi-level Self-supervised Learning for Multimodal Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/33408.

Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4256–4266, June 2025.

Hao Yan, Chaozhuo Li, Zhigang Yu, Jun Yin, Ruochen Liu, Peiyan Zhang, Weihao Han, Mingzheng Li, Zhengxin Zeng, Hao Sun, Weiwei Deng, Feng Sun, Qi Zhang, and Senzhang Wang. When graph meets multimodal: Benchmarking on multimodal attributed graphs learning. *CoRR*, abs/2410.09132, 2024.

Yongyi Yang, Tang Liu, Yangkun Wang, Jinjing Zhou, Quan Gan, Zhewei Wei, Zheng Zhang, Zengfeng Huang, and David Wipf. Graph neural networks inspired by classical iterative algorithms. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 11773–11783, 2021. URL https://proceedings.mlr.press/v139/yang21g.html.

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

Sirui Zhang, Xu Wang, Zhengyu Wu, Xunkai Li, and Hongchao Qin. Context-aware modality-topology co-alignment for multimodal attributed graphs. *arXiv preprint arXiv:2606.14172*, 2026.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging Modalities: Improving Universal Multimodal Retrieval by Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Bridging_Modalities_Improving_Universal_Multimodal_Retrieval_by_Multimodal_Large_Language_CVPR_2025_paper.html.

Haoran Zheng, Renchi Yang, Hongtao Wang, and Jianliang Xu. Cross-contrastive clustering for multimodal attributed graphs with dual graph filtering, 2025.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. VISTA: Visualized Text Embedding For Universal Multi-Modal Retrieval. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://aclanthology.org/2024.acl-long.175/.

Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap Latent Representations for Multi-Modal Recommendation. In *Proceedings of the ACM Web Conference*, 2023. URL https://arxiv.org/abs/2207.05969.

Jing Zhu, Yuhang Zhou, Shengyi Qian, Zhongmou He, Tong Zhao, Neil Shah, and Danai Koutra. Mosaic of modalities: A comprehensive benchmark for multimodal graph learning. In *CVPR*, pp. 14215–14224. Computer Vision Foundation / IEEE, 2025.

Tianyu Zhu, Myong Chol Jung, and Jesse Clark. Generalized Contrastive Learning for Multi-Modal Retrieval and Ranking. *arXiv preprint arXiv:2404.08535*, 2024. URL https://arxiv.org/abs/2404.08535.

Yilong Zuo, Xunkai Li, Zhihan Zhang, Ronghua Li, and Guoren Wang. RoleMAG: Learning neighbor roles in multimodal graphs. *arXiv preprint arXiv:2604.12271*, 2026.

APPENDIX CONTENTS

Appendix A discusses related work. Appendix B gives all four retrieval comparisons and extended analyses. Appendix C derives the graph-signal relations. Appendix D specifies the algorithm and optimization. Appendix E describes the data and evaluation. Appendix F details the external-method adaptations. Appendix G summarizes the study's limitations.

A RELATED WORK

A.1 VISION-LANGUAGE RETRIEVAL

Paired image-text learning establishes a common space in which corresponding attributes can be compared. VSE++ emphasizes difficult incorrect matches through hard-negative ranking, whereas CLIP uses bidirectional contrastive supervision over large collections of paired examples (Faghri et al., 2017; Radford et al., 2021). SigLIP uses pairwise sigmoid learning, and SigLIP 2 extends this family with additional training and data-curation procedures (Zhai et al., 2023; Tschannen et al., 2025). CLIP-Refine studies feature-level modality alignment after pretraining (Yamaguchi et al., 2025). These approaches motivate GOMA's use of paired supervision on supplied features. The graph adds relationships among different objects, making the retrieval target a further consideration when deciding how to organize the learned representation.

Universal multimodal retrieval broadens the possible query and document interfaces. UniIR trains instruction-guided retrievers across modality combinations and introduces the M-BEIR benchmark (Wei et al., 2024). VISTA adds visual information to text embedding models, VLM2Vec trains vision-language models for embedding tasks, and MM-Embed addresses modality bias through modality-aware hard negatives (Zhou et al., 2024; Jiang et al., 2025; Lin et al., 2025a). GME expands multimodal retrieval training through synthesized data, while UniECS studies gated fusion for product search (Zhang et al., 2025; Liang et al., 2025). Generalized contrastive learning further incorporates richer relevance and ranking objectives (Zhu et al., 2024; Lee et al., 2025). This line distinguishes retrieval relevance from the recovery of one paired attribute. GOMA additionally considers observed object relationships and explicitly separates attribute-only queries from dual-attribute retrieval over complete records.

Representations may retain information at several semantic granularities. LAPS selects and refines visual patches for patch-word alignment (Fu et al., 2024). AVSE studies asymmetric visual-semantic embeddings, D2S-VSE transfers information between dense and sparse descriptions, and SmartCLIP uses textual masks to select visual concepts (Liu et al., 2025c;b; Xie et al., 2025). These methods address the unequal content of images and descriptions. GOMA uses object-level feature inputs and distinguishes the desired retrieval outputs through the position of their supervision. It learns image-text correspondence at both embeddings, while graph smoothness acts on the semantic representation used for cross-object relevance.

The distinction between a representation and a projection trained from it is established in contrastive learning. SimCLR shows that the representation before a nonlinear projection can retain information useful for downstream prediction that is less available after projection (Chen et al., 2020). GOMA uses this architectural separation for two jointly supervised retrieval outputs. The content embedding is trained for paired identity, the semantic embedding learns graph-regularized cross-modal correspondence, and both remain available for inference. The shared readout comparison in Table 4 specifically tests this supervision placement at equal parameter count.

Relations also enter retrieval at different levels. GSMN models structure inside an image-text comparison (Liu et al., 2020), and instance-level semantic relationship learning extends supervision across examples (Fu et al., 2023). NeighborRetr addresses frequently retrieved hubs through neighborhood-aware training (Lin et al., 2025b). IN2R uses intramodal neighbors and graph reasoning to refine noisy cross-modal correspondence (Liu et al., 2026). These studies involve local phrase structure, learned neighborhoods, or uncertain training pairs. Our setting supplies observed inter-object edges alongside known paired attributes. The graph regularizes semantic outputs during learning and is used for propagation in dual-attribute retrieval.

A.2 Multimodal Graph Learning

Multimodal graph benchmarks evaluate topology together with visual and textual object attributes. MM-GRAPH studies visual information in graph learning, MAGB organizes multimodal-attributed graph tasks, and OpenMAG broadens graph-level and modality-level evaluation (Zhu et al., 2025; Yan et al., 2024; Wan et al., 2026). Graph-MLLM studies how multimodal language models can supply encoders, aligners, and predictors for graph tasks (Liu et al., 2025a). These resources motivate clear definitions of which attributes, relationships, and labels enter a model. Our experiments use common prepared features and retrieval pools so that each adapted method is evaluated on the same supplied information.

Neighborhood aggregation provides a direct way to incorporate a graph. GCN uses normalized aggregation, GAT learns attention over neighbors, and GraphSAGE develops inductive aggregation functions (Kipf & Welling, 2017; Veličković et al., 2018; Hamilton et al., 2017). Multimodal recommenders introduce modality-specific graph processing and fusion, as in MMGCN and MGAT, or combine local and global relationships, as in LGMRec (Wei et al., 2019; Tao et al., 2020; Guo et al., 2024). MIG-GT separates modality-specific neighborhoods and includes global interactions (Hu et al., 2025). BM3 learns through bootstrapped latent views, and MENTOR uses interaction-derived guidance to align modalities (Zhou et al., 2023; Xu et al., 2025). These models retain signals useful for preference prediction. GOMA instead trains embeddings whose scores correspond to paired identity or cross-object relevance, using observed object relations for the latter.

Recent methods make modality-topology interactions more selective. DMGC distinguishes homophilous and heterophilous information for clustering, and DGF combines graph filtering with cross-contrastive clustering (Guo et al., 2025; Zheng et al., 2025). CAMPA separates propagation from aggregation, CoMAG models context-dependent modality-topology alignment, and RoleMAG learns neighbor roles (Su et al., 2026; Zhang et al., 2026; Zuo et al., 2026). LION introduces Clifford-algebra propagation and adaptive aggregation for multimodal-attributed graphs (Li et al., 2026). These designs change how structural and modality information interact. GOMA uses a normalized graph Dirichlet regularizer on the semantic embedding and keeps identity supervision at the preceding content representation, making retrieval relevance determine which learned output is used.

Multimodal graph pretraining seeks reusable representations across datasets and tasks. UniGraph2 combines modality encoders, alignment, and graph processing across domains (He et al., 2025). Graph4MM incorporates multihop structural information into multimodal learning (Ning et al., 2025). MLaGA and GraphGPT-O connect graphs with multimodal language-model reasoning and generation (Fan et al., 2025; Fang et al., 2025). NTSFormer addresses isolated cold-start nodes and missing modalities with self-teaching and attention masks (Hu et al., 2026). GOMA studies the retrieval interface of a jointly learned model. Independent attribute encoders handle single-modality inputs, and their semantic outputs guide propagation for complete records.

SMGFM is especially close in its distinction between structural consistency and modality-specific detail (Wu et al., 2026). It constructs Chebyshev frequency-band tokens and assigns their roles before modality interaction. GOMA organizes the distinction by retrieval relevance and supervision position. Its content and semantic embeddings are jointly learned mappings, and the observed graph penalizes variation in the semantic output. The same semantic signal also guides the weights of observed edges during propagation. This connects retrieval-specific supervision to neighborhood propagation, with the content embedding remaining directly accessible for identity retrieval.

A.3 Graph Signal Processing and Retrieval Diffusion

Graph signal processing describes feature variation relative to graph topology through Laplacian energy and spectral filters (Shuman et al., 2013). APPNP combines learned predictions with personalized propagation, and SLOG studies inductive spectral filtering beyond polynomial filters (Klicpera et al., 2019; Xu et al., 2024). Reverse message passing and spectral graph pruning modify propagation or topology to mitigate unwanted smoothing (Park et al., 2024; Jamadandi et al., 2024). Recent analyses also distinguish the roles of residual connections, normalization, and gradient behavior in deep graph networks (Scholkemper et al., 2025; Chen et al., 2025; Park & Kim, 2024). GOMA uses these concepts to explain the graph-regularization term on the semantic embedding and the finite graph filter. The learned nonlinear mappings, the linear propagation recurrence, and final row normalization have separate mathematical roles.

Diffusion is an established tool for retrieval. Iscen et al. propagate similarities on a manifold of regional visual descriptors (Iscen et al., 2017). Regularized diffusion studies affinity propagation through higher-order graph relationships (Bai et al., 2017). Matrix Completion on Graphs applies Dirichlet regularization along row and column graphs (Kalofolias et al., 2014). SRFeat combines smoothness and correspondence learning in geometric matching (Li et al., 2022). These works provide precedents for using graph variation to organize matching signals. GOMA connects a graph-regularized semantic embedding to a content embedding for identity retrieval and to propagation over observed object relationships. TWIRLS derives graph aggregation and adaptive edge reweighting from iterative optimization (Yang et al., 2021). GOMA computes continuous edge weights from learned cross-modal semantic agreement, holds them unchanged during propagation, and uses the propagated representations for dual-attribute retrieval. The contribution is this connection between retrieval-specific graph regularization and propagation, alongside independently accessible identity and semantic outputs.

B Additional Experimental Results

B.1 Complete Retrieval Results

The following four tables give the complete external comparison for the task definitions in Sec. 2. Every table contains the same 14 external methods and GOMA. GOMA uses its single selected model for all applicable outputs on a dataset, while external methods retain the task-specific selection described in Appendix F. All repeated-run values are means and sample standard deviations over seeds 43, 44, and 45. The table groups separate vision-language learning, graph encoders, multimodal graph learning, and GOMA.

Table 5 reports identity recall in both directions and their bidirectional mean. It extends the primary R@1 comparison to ranks 5 and 10. The strongest bidirectional R@10 on Reddit-S comes from CLIP-Refine, whereas GOMA leads the bidirectional R@1 values used in the main comparison. Directional results expose the image-to-text and text-to-image differences that an average alone can conceal.

Table 5: Complete single-modality identity retrieval. Directional recalls and their bidirectional mean use the same models.

Bidirectional mean						
Method	**Grocery**			**Reddit-S**		
	R@1	R@5	R@10	R@1	R@5	R@10
Vision-language learning						
VSE++	49.94 ± 0.10	75.53 ± 0.38	82.26 ± 0.26	16.10 ± 0.40	33.98 ± 0.27	42.92 ± 0.20
CLIP-Refine	49.93 ± 0.07	76.91 ± 0.17	84.02 ± 0.09	<u>18.48</u> ± 0.11	<u>37.91</u> ± 0.17	**47.47** ± 0.02
SigLIP	<u>52.36</u> ± 0.30	<u>78.05</u> ± 0.17	<u>85.12</u> ± 0.05	17.14 ± 0.19	35.89 ± 0.32	45.18 ± 0.30
Graph encoders						
GCN	4.60 ± 0.13	13.48 ± 0.68	19.70 ± 0.72	0.97 ± 0.24	3.56 ± 0.20	5.92 ± 0.13
GAT	5.97 ± 1.07	15.76 ± 1.53	22.45 ± 1.43	1.20 ± 0.17	4.60 ± 0.24	7.28 ± 0.45
GraphSAGE	1.82 ± 0.38	6.14 ± 1.30	9.73 ± 2.07	0.56 ± 0.16	1.99 ± 0.36	3.24 ± 0.49
Multimodal graph learning						
MMGCN	15.33 ± 1.13	37.46 ± 1.34	49.32 ± 1.44	4.42 ± 0.24	13.05 ± 0.43	19.88 ± 0.54
MGAT	20.82 ± 0.29	45.53 ± 0.97	57.13 ± 1.07	5.92 ± 0.28	17.16 ± 0.84	24.95 ± 0.71
LGMRec	9.39 ± 0.60	24.65 ± 1.07	33.80 ± 1.21	2.16 ± 0.41	7.32 ± 0.49	11.38 ± 0.59
DMGC	24.05 ± 0.95	50.78 ± 1.94	62.91 ± 1.43	6.88 ± 0.31	18.91 ± 0.56	27.48 ± 0.21
DGF	20.29 ± 0.62	46.31 ± 0.90	58.47 ± 0.77	5.71 ± 0.48	16.82 ± 0.19	24.80 ± 0.12
CAMPA	3.25 ± 1.15	9.57 ± 2.98	14.25 ± 3.79	0.73 ± 0.14	2.69 ± 0.46	4.34 ± 0.81
CoMAG	12.15 ± 0.87	29.08 ± 1.36	38.14 ± 1.72	2.34 ± 0.34	7.26 ± 0.32	10.72 ± 0.81
RoleMAG	5.57 ± 0.25	15.54 ± 0.59	22.14 ± 0.75	0.98 ± 0.11	3.60 ± 0.17	5.71 ± 0.27
Proposed method						
GOMA	**52.96** ± 0.06	**78.57** ± 0.06	**85.14** ± 0.17	**18.66** ± 0.19	**38.13** ± 0.26	<u>47.37</u> ± 0.36

Continued on the next page

Table 5 continued

Bidirectional mean

Method	Toys			Books		
	R@1	R@5	R@10	R@1	R@5	R@10
Vision-language learning						
VSE++	59.19 ± 0.12	82.20 ± 0.10	87.49 ± 0.08	39.23 ± 0.09	61.23 ± 0.09	68.98 ± 0.03
CLIP-Refine	62.57 ± 0.10	85.34 ± 0.05	90.22 ± 0.04	35.44 ± 0.01	56.33 ± 0.02	64.15 ± 0.04
SigLIP	61.29 ± 0.49	84.78 ± 0.21	90.12 ± 0.18	37.84 ± 0.07	59.87 ± 0.14	67.76 ± 0.08
Graph encoders						
GCN	2.87 ± 0.31	8.48 ± 0.39	12.54 ± 0.41	0.32 ± 0.04	1.15 ± 0.02	1.95 ± 0.05
GAT	4.64 ± 0.78	12.76 ± 0.95	18.18 ± 0.93	0.91 ± 0.19	3.26 ± 0.42	5.30 ± 0.66
GraphSAGE	1.17 ± 0.21	4.00 ± 0.67	6.57 ± 1.10	0.38 ± 0.07	1.52 ± 0.15	2.64 ± 0.23
Multimodal graph learning						
MMGCN	14.26 ± 0.97	33.03 ± 1.95	42.32 ± 2.22	12.44 ± 0.32	28.36 ± 0.46	37.18 ± 0.54
MGAT	22.81 ± 0.73	45.53 ± 0.64	55.90 ± 0.56	17.23 ± 0.54	35.48 ± 0.90	44.48 ± 1.14
LGMRec	8.61 ± 0.11	21.17 ± 0.15	28.33 ± 0.10	1.54 ± 0.02	5.12 ± 0.02	8.05 ± 0.05
DMGC	23.74 ± 0.68	46.28 ± 0.35	55.41 ± 0.41	10.36 ± 0.47	24.02 ± 0.58	31.87 ± 0.75
DGF	20.81 ± 0.41	44.26 ± 0.79	55.18 ± 0.78	13.11 ± 0.52	29.35 ± 0.73	38.31 ± 0.71
CAMPA	3.69 ± 0.75	10.42 ± 1.72	14.78 ± 2.33	0.24 ± 0.06	0.87 ± 0.23	1.47 ± 0.34
CoMAG	11.20 ± 0.91	25.83 ± 1.04	33.85 ± 0.85	1.46 ± 0.16	4.74 ± 0.40	7.35 ± 0.50
RoleMAG	4.07 ± 0.89	10.67 ± 1.22	15.65 ± 1.36	0.61 ± 0.06	2.36 ± 0.25	3.84 ± 0.30
Proposed method						
GOMA	**62.94** ± 0.40	**86.09** ± 0.32	**90.96** ± 0.20	**40.14** ± 0.31	**62.22** ± 0.42	**69.84** ± 0.39

Image to text

Method	Grocery			Reddit-S		
	R@1	R@5	R@10	R@1	R@5	R@10
Vision-language learning						
VSE++	49.80 ± 0.28	75.67 ± 0.30	82.12 ± 0.17	16.21 ± 0.46	34.30 ± 0.05	43.28 ± 0.14
CLIP-Refine	50.05 ± 0.08	76.86 ± 0.20	84.22 ± 0.27	**18.63** ± 0.08	37.86 ± 0.26	47.15 ± 0.10
SigLIP	52.53 ± 0.49	77.98 ± 0.09	**85.14** ± 0.08	17.27 ± 0.26	35.86 ± 0.71	45.27 ± 0.49
Graph encoders						
GCN	4.58 ± 0.34	13.64 ± 0.78	19.89 ± 0.79	0.99 ± 0.38	3.59 ± 0.16	6.13 ± 0.13
GAT	6.05 ± 1.14	15.65 ± 1.63	22.50 ± 1.33	1.15 ± 0.08	4.41 ± 0.28	7.08 ± 0.56
GraphSAGE	1.89 ± 0.54	6.13 ± 1.41	9.73 ± 2.23	0.48 ± 0.05	1.82 ± 0.16	3.02 ± 0.45
Multimodal graph learning						
MMGCN	15.24 ± 1.00	36.94 ± 1.00	48.96 ± 1.70	4.52 ± 0.26	13.40 ± 0.32	20.23 ± 0.17
MGAT	20.13 ± 0.31	45.39 ± 0.84	56.90 ± 1.15	6.14 ± 0.21	17.42 ± 0.91	25.23 ± 1.03
LGMRec	9.21 ± 0.76	24.35 ± 1.32	33.49 ± 1.69	2.06 ± 0.28	6.91 ± 0.41	11.03 ± 0.54
DMGC	23.57 ± 1.01	50.20 ± 1.83	62.75 ± 1.12	6.97 ± 0.36	19.29 ± 0.25	27.79 ± 0.19
DGF	19.94 ± 0.68	45.82 ± 1.26	58.24 ± 1.06	5.91 ± 0.62	17.15 ± 0.23	25.18 ± 0.11
CAMPA	3.45 ± 2.93	10.55 ± 7.78	15.44 ± 10.28	0.43 ± 0.18	1.97 ± 0.54	3.30 ± 0.94
CoMAG	12.69 ± 0.37	30.21 ± 1.28	39.33 ± 1.18	2.47 ± 0.41	6.99 ± 0.35	10.32 ± 0.79
RoleMAG	5.35 ± 0.13	15.51 ± 0.71	21.83 ± 0.74	0.89 ± 0.12	3.59 ± 0.07	5.80 ± 0.05
Proposed method						
GOMA	**52.85** ± 0.06	**78.65** ± 0.22	**85.14** ± 0.15	18.51 ± 0.18	**38.36** ± 0.30	**47.79** ± 0.30

Continued on the next page

Table 5 continued

Image to text Method	Toys			Books		
	R@1	R@5	R@10	R@1	R@5	R@10
Vision-language learning						
VSE++	59.21 ± 0.17	81.84 ± 0.12	87.20 ± 0.03	40.02 ± 0.13	61.79 ± 0.04	69.38 ± 0.12
CLIP-Refine	62.70 ± 0.18	85.05 ± 0.07	90.08 ± 0.04	36.75 ± 0.05	57.45 ± 0.03	65.17 ± 0.03
SigLIP	61.52 ± 0.29	84.34 ± 0.23	89.59 ± 0.23	38.82 ± 0.11	60.62 ± 0.18	68.17 ± 0.10
Graph encoders						
GCN	2.77 ± 0.39	8.25 ± 0.40	12.35 ± 0.35	0.37 ± 0.04	1.24 ± 0.04	2.08 ± 0.07
GAT	4.63 ± 0.84	12.59 ± 1.13	18.02 ± 0.83	0.98 ± 0.16	3.48 ± 0.36	5.60 ± 0.64
GraphSAGE	1.23 ± 0.16	3.97 ± 0.80	6.48 ± 1.21	0.42 ± 0.06	1.64 ± 0.18	2.79 ± 0.21
Multimodal graph learning						
MMGCN	14.19 ± 1.08	33.00 ± 1.70	42.14 ± 2.10	12.60 ± 0.23	28.45 ± 0.39	37.38 ± 0.65
MGAT	23.16 ± 0.57	45.66 ± 0.56	56.13 ± 0.44	17.43 ± 0.59	35.67 ± 0.85	44.65 ± 0.97
LGMRec	8.51 ± 0.17	21.11 ± 0.07	28.19 ± 0.19	1.67 ± 0.09	5.49 ± 0.03	8.59 ± 0.14
DMGC	23.72 ± 0.84	46.04 ± 0.13	55.16 ± 0.19	10.38 ± 0.56	24.16 ± 0.71	32.03 ± 0.81
DGF	21.04 ± 0.45	44.16 ± 0.99	55.05 ± 0.91	13.16 ± 0.51	29.50 ± 0.81	38.57 ± 0.78
CAMPA	3.99 ± 1.64	11.55 ± 4.24	16.22 ± 5.57	0.28 ± 0.03	1.01 ± 0.14	1.71 ± 0.23
CoMAG	11.77 ± 1.01	26.82 ± 0.95	34.92 ± 0.86	1.46 ± 0.23	4.75 ± 0.55	7.31 ± 0.72
RoleMAG	3.93 ± 0.95	10.51 ± 1.15	15.41 ± 1.30	0.65 ± 0.04	2.43 ± 0.21	3.95 ± 0.24
Proposed method						
GOMA	**63.11** ± 0.45	**85.60** ± 0.42	**90.57** ± 0.31	**41.13** ± 0.35	**62.93** ± 0.38	**70.40** ± 0.42

Text to image Method	Grocery			Reddit-S		
	R@1	R@5	R@10	R@1	R@5	R@10
Vision-language learning						
VSE++	50.08 ± 0.25	75.39 ± 0.50	82.40 ± 0.39	15.98 ± 0.37	33.66 ± 0.50	42.56 ± 0.38
CLIP-Refine	49.81 ± 0.10	76.97 ± 0.21	83.81 ± 0.11	18.33 ± 0.23	**37.95** ± 0.13	**47.79** ± 0.08
SigLIP	52.20 ± 0.16	78.11 ± 0.31	85.11 ± 0.06	17.01 ± 0.25	35.92 ± 0.23	45.09 ± 0.12
Graph encoders						
GCN	4.61 ± 0.08	13.32 ± 0.62	19.52 ± 0.79	0.96 ± 0.17	3.52 ± 0.26	5.71 ± 0.18
GAT	5.89 ± 1.08	15.86 ± 1.42	22.40 ± 1.55	1.25 ± 0.28	4.79 ± 0.21	7.48 ± 0.33
GraphSAGE	1.75 ± 0.23	6.16 ± 1.20	9.72 ± 1.93	0.64 ± 0.27	2.16 ± 0.55	3.45 ± 0.53
Multimodal graph learning						
MMGCN	15.42 ± 1.36	37.98 ± 1.69	49.68 ± 1.27	4.31 ± 0.22	12.70 ± 0.60	19.52 ± 0.92
MGAT	21.50 ± 0.50	45.67 ± 1.10	57.35 ± 1.06	5.69 ± 0.46	16.89 ± 0.77	24.67 ± 0.54
LGMRec	9.56 ± 0.62	24.95 ± 0.84	34.12 ± 0.75	2.26 ± 0.57	7.73 ± 0.58	11.74 ± 0.71
DMGC	24.53 ± 0.88	51.37 ± 2.11	63.06 ± 1.75	6.79 ± 0.56	18.54 ± 0.93	27.17 ± 0.54
DGF	20.65 ± 0.59	46.80 ± 0.60	58.71 ± 0.54	5.50 ± 0.37	16.49 ± 0.15	24.42 ± 0.28
CAMPA	3.06 ± 0.79	8.60 ± 2.16	13.06 ± 2.96	1.02 ± 0.37	3.42 ± 1.22	5.38 ± 1.93
CoMAG	11.61 ± 1.38	27.95 ± 1.44	36.95 ± 2.33	2.20 ± 0.28	7.53 ± 0.70	11.12 ± 1.03
RoleMAG	5.79 ± 0.59	15.57 ± 0.55	22.44 ± 0.81	1.07 ± 0.16	3.61 ± 0.29	5.63 ± 0.52
Proposed method						
GOMA	**53.07** ± 0.15	**78.50** ± 0.16	**85.14** ± 0.21	**18.80** ± 0.25	37.90 ± 0.24	46.96 ± 0.42

Continued on the next page

Table 5 continued

Method	Toys			Books		
Text to image						
	R@1	R@5	R@10	R@1	R@5	R@10
Vision-language learning						
VSE++	59.17 ± 0.06	82.57 ± 0.09	87.78 ± 0.14	38.44 ± 0.14	60.67 ± 0.17	68.58 ± 0.08
CLIP-Refine	62.44 ± 0.11	85.63 ± 0.16	90.36 ± 0.04	34.14 ± 0.03	55.20 ± 0.04	63.14 ± 0.06
SigLIP	61.05 ± 0.69	85.21 ± 0.19	90.65 ± 0.13	36.88 ± 0.23	59.13 ± 0.11	67.36 ± 0.12
Graph encoders						
GCN	2.97 ± 0.25	8.70 ± 0.39	12.74 ± 0.51	0.26 ± 0.03	1.05 ± 0.01	1.83 ± 0.05
GAT	4.64 ± 0.72	12.93 ± 0.79	18.34 ± 1.05	0.85 ± 0.22	3.03 ± 0.51	5.00 ± 0.71
GraphSAGE	1.12 ± 0.27	4.02 ± 0.54	6.66 ± 1.02	0.35 ± 0.09	1.40 ± 0.14	2.50 ± 0.24
Multimodal graph learning						
MMGCN	14.33 ± 0.87	33.06 ± 2.21	42.50 ± 2.35	12.26 ± 0.43	28.27 ± 0.55	36.98 ± 0.52
MGAT	22.46 ± 0.88	45.41 ± 0.81	55.68 ± 0.69	17.03 ± 0.48	35.29 ± 0.97	44.30 ± 1.31
LGMRec	8.70 ± 0.37	21.24 ± 0.35	28.47 ± 0.01	1.42 ± 0.09	4.75 ± 0.05	7.50 ± 0.04
DMGC	23.76 ± 0.57	46.51 ± 0.74	55.67 ± 0.65	10.33 ± 0.38	23.87 ± 0.44	31.72 ± 0.69
DGF	20.58 ± 0.39	44.36 ± 0.59	55.31 ± 0.68	13.07 ± 0.54	29.19 ± 0.67	38.05 ± 0.65
CAMPA	3.40 ± 1.92	9.30 ± 4.02	13.33 ± 5.68	0.19 ± 0.09	0.73 ± 0.32	1.23 ± 0.51
CoMAG	10.62 ± 0.91	24.84 ± 1.13	32.78 ± 0.86	1.46 ± 0.20	4.74 ± 0.43	7.40 ± 0.66
RoleMAG	4.22 ± 0.84	10.83 ± 1.30	15.89 ± 1.41	0.57 ± 0.09	2.30 ± 0.31	3.73 ± 0.35
Proposed method						
GOMA	**62.77** ± 0.36	**86.58** ± 0.27	**91.35** ± 0.09	**39.15** ± 0.27	**61.51** ± 0.46	**69.29** ± 0.36

Table 6 extends single-modality semantic retrieval to precision and semantic recall. GOMA has the highest mean for every displayed metric on all four datasets. The query and gallery remain disjoint, and each attribute is encoded independently, so the results concern cross-object semantic ranking from a single modality.

Table 6: Complete single-modality cross-object semantic retrieval. R@1 uses semantic relevance.

Method	Grocery			Reddit-S		
	mAP@10	P@10	Sem. R@1	mAP@10	P@10	Sem. R@1
Vision-language learning						
VSE++	30.53 ± 0.14	39.26 ± 0.10	55.54 ± 0.32	37.34 ± 0.26	46.21 ± 0.28	59.52 ± 0.11
CLIP-Refine	33.72 ± 0.06	42.18 ± 0.06	58.66 ± 0.10	46.19 ± 0.10	54.02 ± 0.09	64.65 ± 0.14
SigLIP	38.27 ± 0.12	46.31 ± 0.10	59.52 ± 0.13	50.89 ± 0.04	58.29 ± 0.08	66.59 ± 0.42
Graph encoders						
GCN	8.86 ± 0.53	15.73 ± 0.62	19.54 ± 1.22	11.16 ± 1.20	18.76 ± 1.32	22.05 ± 1.55
GAT	11.96 ± 0.42	19.16 ± 0.46	23.43 ± 0.75	16.41 ± 1.46	24.65 ± 1.78	28.87 ± 2.21
GraphSAGE	6.78 ± 0.45	12.81 ± 0.64	14.60 ± 1.34	10.17 ± 1.26	16.96 ± 1.52	18.33 ± 1.45
Multimodal graph learning						
MMGCN	28.15 ± 0.53	36.57 ± 0.61	45.53 ± 1.42	42.22 ± 0.36	50.13 ± 0.37	55.40 ± 0.30
MGAT	30.54 ± 1.00	38.66 ± 0.98	49.62 ± 1.23	43.60 ± 0.98	51.39 ± 0.72	58.49 ± 0.95
LGMRec	12.88 ± 0.23	20.83 ± 0.36	26.18 ± 0.58	18.79 ± 0.39	27.15 ± 0.30	31.45 ± 1.02
DMGC	34.70 ± 0.23	42.67 ± 0.27	53.45 ± 0.39	50.60 ± 0.27	57.26 ± 0.23	61.56 ± 0.47
DGF	32.45 ± 0.96	40.58 ± 0.86	50.41 ± 0.57	44.62 ± 0.97	52.38 ± 0.95	58.20 ± 0.54
CAMPA	8.76 ± 0.78	14.93 ± 1.19	18.23 ± 1.65	10.52 ± 1.77	15.78 ± 1.96	17.75 ± 1.56
CoMAG	13.84 ± 0.75	21.69 ± 0.70	29.38 ± 1.72	17.05 ± 1.51	25.12 ± 1.76	30.38 ± 1.73
RoleMAG	9.32 ± 0.31	16.51 ± 0.44	20.87 ± 0.80	13.06 ± 0.65	20.72 ± 0.21	24.95 ± 0.74
Proposed method						
GOMA	**42.82** ± 0.22	**50.23** ± 0.17	**60.16** ± 0.39	**59.74** ± 0.26	**65.39** ± 0.31	**70.13** ± 0.08

Continued on the next page

Table 6 continued

Method	Toys			Books		
	mAP@10	P@10	Sem. R@1	mAP@10	P@10	Sem. R@1
Vision-language learning						
VSE++	23.09 ± 0.09	32.35 ± 0.09	50.78 ± 0.46	39.99 ± 0.29	50.51 ± 0.23	55.66 ± 0.28
CLIP-Refine	27.57 ± 0.05	37.01 ± 0.08	54.50 ± 0.04	42.82 ± 0.02	52.98 ± 0.02	57.32 ± 0.06
SigLIP	32.75 ± 0.25	42.04 ± 0.33	57.46 ± 0.02	42.42 ± 0.07	52.57 ± 0.01	57.10 ± 0.10
Graph encoders						
GCN	4.55 ± 0.06	10.26 ± 0.07	11.43 ± 0.60	16.65 ± 0.99	26.78 ± 1.08	27.87 ± 1.25
GAT	5.08 ± 0.49	11.08 ± 0.78	12.71 ± 1.20	23.08 ± 0.67	33.81 ± 0.93	35.45 ± 0.84
GraphSAGE	3.46 ± 0.17	8.55 ± 0.38	8.85 ± 0.60	17.29 ± 1.20	27.67 ± 1.39	29.05 ± 1.31
Multimodal graph learning						
MMGCN	16.94 ± 0.76	25.50 ± 0.73	35.20 ± 1.25	40.82 ± 0.73	50.91 ± 0.66	53.91 ± 0.55
MGAT	20.51 ± 0.31	29.17 ± 0.22	41.13 ± 0.15	41.71 ± 0.12	51.61 ± 0.09	55.16 ± 0.17
LGMRec	6.50 ± 0.43	13.26 ± 0.57	16.64 ± 0.40	25.63 ± 0.98	36.54 ± 0.98	38.27 ± 1.10
DMGC	17.48 ± 0.35	26.10 ± 0.43	36.02 ± 0.58	42.14 ± 0.37	52.03 ± 0.30	54.66 ± 0.48
DGF	23.01 ± 0.64	31.86 ± 0.37	42.65 ± 0.48	40.89 ± 0.27	50.99 ± 0.24	54.31 ± 0.16
CAMPA	5.37 ± 0.52	11.03 ± 0.79	13.27 ± 1.10	16.07 ± 1.49	24.84 ± 2.53	26.32 ± 2.75
CoMAG	7.18 ± 0.22	14.20 ± 0.25	19.68 ± 0.35	26.37 ± 0.61	37.11 ± 0.53	38.94 ± 0.31
RoleMAG	4.93 ± 0.33	10.77 ± 0.65	12.93 ± 0.84	20.76 ± 0.85	31.21 ± 0.78	32.67 ± 0.89
Proposed method						
GOMA	**41.52** ± 0.44	**50.22** ± 0.48	**61.36** ± 0.30	**46.16** ± 0.03	**55.79** ± 0.01	**59.79** ± 0.07

Table 7 reports the same semantic metrics when complete object records and the permitted observed graph are available. The relevance definition is unchanged from Table 6, allowing the two interfaces to be compared without confusing a semantic match with paired identity. GOMA leads all displayed means in this dual-attribute comparison.

Table 7: Complete dual-attribute cross-object semantic retrieval. R@1 uses semantic relevance.

Method	Grocery			Reddit-S		
	mAP@10	P@10	Sem. R@1	mAP@10	P@10	Sem. R@1
Vision-language learning						
VSE++	39.67 ± 0.13	47.82 ± 0.14	63.75 ± 0.37	56.19 ± 0.39	62.99 ± 0.27	76.16 ± 0.46
CLIP-Refine	34.81 ± 0.32	42.92 ± 0.37	60.87 ± 0.57	59.37 ± 2.67	65.76 ± 2.79	75.77 ± 0.35
SigLIP	44.17 ± 0.08	52.06 ± 0.11	66.15 ± 0.28	62.95 ± 0.11	69.05 ± 0.14	77.85 ± 0.34
Graph encoders						
GCN	36.50 ± 1.06	43.39 ± 0.97	54.02 ± 0.92	64.61 ± 0.86	69.61 ± 0.77	77.65 ± 0.44
GAT	35.67 ± 0.64	42.49 ± 0.47	52.04 ± 1.38	67.93 ± 0.67	72.45 ± 0.36	80.46 ± 0.80
GraphSAGE	28.95 ± 1.89	35.41 ± 2.25	49.02 ± 1.46	58.01 ± 1.22	63.20 ± 0.93	74.59 ± 1.12
Multimodal graph learning						
MMGCN	35.21 ± 0.82	42.00 ± 0.94	52.56 ± 0.82	64.74 ± 1.29	69.92 ± 1.07	76.55 ± 1.09
MGAT	40.65 ± 0.75	48.06 ± 0.82	61.40 ± 0.25	67.43 ± 1.69	72.77 ± 1.39	78.66 ± 0.72
LGMRec	41.80 ± 1.64	48.86 ± 1.79	63.69 ± 1.50	69.29 ± 0.44	73.61 ± 0.37	81.66 ± 0.88
DMGC	32.56 ± 11.45	40.21 ± 11.74	58.96 ± 9.09	55.13 ± 1.47	62.08 ± 1.31	73.95 ± 0.32
DGF	34.57 ± 0.54	41.73 ± 0.58	60.45 ± 1.57	61.93 ± 0.14	67.66 ± 0.13	78.49 ± 0.64
CAMPA	30.12 ± 1.15	38.17 ± 1.14	55.33 ± 0.96	56.32 ± 0.59	62.63 ± 0.30	74.17 ± 0.32
CoMAG	40.31 ± 1.91	47.43 ± 1.92	62.88 ± 0.32	67.28 ± 1.53	71.72 ± 1.59	81.17 ± 1.30
RoleMAG	38.11 ± 0.32	44.73 ± 0.30	57.43 ± 0.77	65.43 ± 1.21	70.36 ± 1.07	78.79 ± 0.41
Proposed method						
GOMA	**49.60** ± 0.05	**56.13** ± 0.03	**67.82** ± 0.15	**78.97** ± 0.05	**81.80** ± 0.03	**85.26** ± 0.14

Continued on the next page

Table 7 continued

Method	Toys			Books		
	mAP@10	P@10	Sem. R@1	mAP@10	P@10	Sem. R@1
Vision-language learning						
VSE++	29.58 ± 0.64	38.65 ± 0.73	58.56 ± 0.36	46.23 ± 0.43	55.89 ± 0.37	63.17 ± 0.52
CLIP-Refine	25.79 ± 0.76	34.62 ± 0.74	55.76 ± 0.45	49.98 ± 0.02	59.35 ± 0.02	65.60 ± 0.08
SigLIP	38.44 ± 0.10	47.70 ± 0.11	62.79 ± 0.08	48.84 ± 0.13	58.35 ± 0.15	64.73 ± 0.04
Graph encoders						
GCN	36.15 ± 0.93	42.85 ± 0.95	60.64 ± 0.85	62.90 ± 0.21	69.83 ± 0.19	75.16 ± 0.21
GAT	37.76 ± 0.99	44.65 ± 1.28	61.30 ± 0.26	62.03 ± 0.23	69.00 ± 0.19	74.21 ± 0.41
GraphSAGE	25.90 ± 1.64	32.37 ± 1.92	54.12 ± 0.64	61.06 ± 0.40	68.38 ± 0.32	74.26 ± 0.31
Multimodal graph learning						
MMGCN	26.84 ± 0.65	34.19 ± 0.72	53.17 ± 0.93	61.14 ± 0.57	68.40 ± 0.47	73.65 ± 0.30
MGAT	29.65 ± 0.04	37.69 ± 0.09	56.88 ± 0.37	60.58 ± 0.43	68.05 ± 0.41	73.44 ± 0.07
LGMRec	36.89 ± 0.35	44.05 ± 0.45	62.94 ± 0.92	62.99 ± 0.40	70.32 ± 0.35	75.68 ± 0.33
DMGC	13.53 ± 0.58	20.38 ± 0.73	43.00 ± 1.42	48.56 ± 0.23	57.91 ± 0.23	64.36 ± 0.04
DGF	27.34 ± 3.98	35.01 ± 4.08	56.55 ± 0.64	59.93 ± 0.32	67.85 ± 0.32	72.55 ± 0.15
CAMPA	24.53 ± 1.37	33.08 ± 1.34	54.35 ± 1.48	47.21 ± 1.07	56.57 ± 0.72	62.40 ± 0.68
CoMAG	34.36 ± 2.69	41.85 ± 2.58	62.10 ± 1.26	59.38 ± 1.28	67.06 ± 1.12	73.05 ± 0.89
RoleMAG	36.35 ± 1.02	43.13 ± 1.01	61.43 ± 0.64	62.26 ± 0.28	69.43 ± 0.21	74.93 ± 0.18
Proposed method						
GOMA	**51.31** ± 0.07	**58.87** ± 0.07	**67.72** ± 0.18	**66.70** ± 0.01	**73.19** ± 0.01	**78.12** ± 0.03

Table 8 reports relation MRR and Hits@1/10. A candidate is correct when it is the held-out linked object, and the target relationship is absent from the permitted observed graph. GOMA's leading means extend from reciprocal rank to both hit rates, supporting the use of the learned semantic initialization and observed neighborhoods for missing-relation ranking.

Table 8: Complete dual-attribute relation ranking.

Method	Sports			Cloth		
	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10
Vision-language learning						
VSE++	62.61 ± 0.77	49.97 ± 0.94	86.33 ± 0.38	59.20 ± 0.87	46.78 ± 1.00	83.78 ± 0.56
CLIP-Refine	67.15 ± 0.02	55.10 ± 0.05	89.34 ± 0.01	62.29 ± 0.02	49.86 ± 0.03	86.17 ± 0.03
SigLIP	67.09 ± 0.16	54.99 ± 0.20	89.42 ± 0.08	61.36 ± 1.64	49.01 ± 1.52	85.75 ± 1.39
Graph encoders						
GCN	87.50 ± 0.18	80.68 ± 0.18	98.11 ± 0.23	80.34 ± 0.11	71.81 ± 0.15	94.72 ± 0.15
GAT	86.67 ± 0.07	79.67 ± 0.11	97.76 ± 0.04	79.48 ± 0.12	70.89 ± 0.18	93.99 ± 0.22
GraphSAGE	82.05 ± 0.42	73.44 ± 0.76	95.95 ± 0.45	75.33 ± 0.03	65.76 ± 0.02	92.17 ± 0.39
Multimodal graph learning						
MMGCN	75.48 ± 0.43	64.35 ± 0.60	94.72 ± 0.10	67.29 ± 0.29	55.53 ± 0.51	89.23 ± 0.54
MGAT	73.58 ± 0.41	61.53 ± 0.65	95.07 ± 0.33	66.23 ± 0.45	53.73 ± 0.55	89.66 ± 0.33
LGMRec	87.07 ± 0.11	79.92 ± 0.24	98.34 ± 0.02	80.54 ± 0.32	71.64 ± 0.48	95.57 ± 0.11
DMGC	57.38 ± 16.31	46.07 ± 15.04	78.32 ± 18.33	54.10 ± 13.77	42.57 ± 12.70	76.83 ± 15.65
DGF	77.06 ± 0.37	66.60 ± 0.31	95.03 ± 0.42	71.91 ± 0.85	60.91 ± 1.13	91.79 ± 0.37
CAMPA	64.07 ± 0.34	51.43 ± 0.57	87.75 ± 0.36	56.14 ± 2.12	43.59 ± 2.37	80.71 ± 1.33
CoMAG	84.54 ± 0.64	76.77 ± 0.86	97.26 ± 0.27	74.33 ± 2.55	64.21 ± 3.10	92.43 ± 1.15
RoleMAG	85.20 ± 0.83	77.19 ± 1.14	97.90 ± 0.15	76.83 ± 0.52	66.96 ± 0.70	93.93 ± 0.18
Proposed method						
GOMA	**87.89** ± 0.03	**80.98** ± 0.04	**98.68** ± 0.03	**81.32** ± 0.11	**72.75** ± 0.13	**95.70** ± 0.10

B.2 Component Ablations

The controls distinguish removing a computation from changing its role. The w/o graph regularization variant removes $\mathcal{E}_{\mathcal{G}}$ while retaining both mappings, both paired objectives, and semantic-guided propagation. The w/o semantic mapping variant sets $\mathbf{z} = \mathbf{c}$ and removes g_m. Shared readout keeps all parameters but moves identity supervision and scoring to $\mathbf{z}$. The w/o graph propagation variant scores the combined initialization $\mathbf{H}_0$ directly. These variants are retrained with their own validation scores and the same selection rule as GOMA. Active loss scales retain their full-model initialization values.

Relabeled training graph applies a vertex permutation to the training adjacency. It preserves the graph structure and degree multiset while changing their assignment to objects. The initialization controls remove either $\mathbf{C}$ or $\mathbf{Z}$ from $\mathbf{H}_0$, and the content-guided variant computes edge weights from $\mathbf{C}$. All other operations remain active. These additional controls test the interaction of learned signals with graph propagation.

Tables 9 and 10 give the single-modality results. GOMA leads identity R@1 and semantic mAP@10 over both the same-capacity shared readout and the mapping removal on every category dataset. Graph regularization improves semantic mAP@10 over its removal on all four datasets. Relabeling the training graph also lowers this metric. These comparisons identify useful information from the observed relationships beyond the two paired objectives.

Table 9: Complete bidirectional identity recalls for retrained representation, regularization, and propagation controls.

Variant	Grocery			Reddit-S		
	R@1	R@5	R@10	R@1	R@5	R@10
GOMA	**52.96** ± 0.06	78.57 ± 0.06	85.14 ± 0.17	18.66 ± 0.19	38.13 ± 0.26	47.37 ± 0.36
w/o graph regularization	52.79 ± 0.08	78.43 ± 0.30	84.93 ± 0.50	**18.96** ± 0.55	38.29 ± 0.42	47.34 ± 0.24
w/o graph propagation	52.87 ± 0.09	**78.87** ± 0.28	85.41 ± 0.24	18.70 ± 0.14	38.17 ± 0.20	47.55 ± 0.09
Shared readout	51.79 ± 0.13	78.47 ± 0.22	85.36 ± 0.39	18.13 ± 0.30	37.26 ± 0.51	47.31 ± 0.35
w/o semantic mapping	51.83 ± 0.12	78.77 ± 0.06	**85.50** ± 0.11	18.21 ± 0.31	37.41 ± 0.35	47.46 ± 0.51
Relabeled training graph	51.80 ± 0.15	78.22 ± 0.13	85.07 ± 0.09	18.47 ± 0.09	37.96 ± 0.54	46.80 ± 0.70
w/o content initialization	**52.96** ± 0.06	78.57 ± 0.06	85.14 ± 0.17	18.62 ± 0.24	38.10 ± 0.32	47.39 ± 0.33
w/o semantic initialization	52.87 ± 0.09	**78.87** ± 0.28	85.41 ± 0.24	18.76 ± 0.12	**38.43** ± 0.29	**47.72** ± 0.26
Content-guided edges	**52.96** ± 0.06	78.57 ± 0.06	85.14 ± 0.17	18.70 ± 0.14	38.17 ± 0.20	47.55 ± 0.09
Variant	Toys			Books		
	R@1	R@5	R@10	R@1	R@5	R@10
GOMA	62.94 ± 0.40	86.09 ± 0.32	90.96 ± 0.20	40.14 ± 0.31	62.22 ± 0.42	69.84 ± 0.39
w/o graph regularization	62.51 ± 0.08	85.57 ± 0.17	90.56 ± 0.11	**40.38** ± 0.12	**62.32** ± 0.18	**70.04** ± 0.16
w/o graph propagation	62.52 ± 0.33	85.87 ± 0.32	90.86 ± 0.30	40.04 ± 0.37	62.07 ± 0.39	69.73 ± 0.39
Shared readout	60.89 ± 0.23	84.64 ± 0.04	90.15 ± 0.13	39.00 ± 0.10	60.75 ± 0.06	68.30 ± 0.08
w/o semantic mapping	60.86 ± 0.17	84.65 ± 0.23	90.17 ± 0.21	39.04 ± 0.07	60.78 ± 0.02	68.40 ± 0.08
Relabeled training graph	62.10 ± 0.60	85.15 ± 0.60	90.17 ± 0.46	40.02 ± 0.16	62.09 ± 0.21	69.74 ± 0.30
w/o content initialization	62.87 ± 0.38	86.07 ± 0.30	91.00 ± 0.24	39.96 ± 0.23	62.02 ± 0.32	69.70 ± 0.34
w/o semantic initialization	**62.96** ± 0.26	**86.20** ± 0.06	**91.23** ± 0.17	39.83 ± 0.45	61.89 ± 0.65	69.56 ± 0.61
Content-guided edges	**62.96** ± 0.36	86.14 ± 0.25	91.09 ± 0.03	40.19 ± 0.36	62.30 ± 0.46	69.95 ± 0.44

Table 11 reports every dual-attribute metric. Full GOMA leads both core graph removals on five datasets. On Cloth, graph-regularization removal increases MRR by 0.131 points on average. The content-initialization removal improves some Reddit-S and Toys scores, and semantic guidance is not the strongest propagation rule on the two relation datasets. These results distinguish the roles of graph regularization and graph propagation across retrieval targets. The same GOMA configuration is used throughout.

Fig. 6 gives paired uncertainty for the core ablations. GOMA's primary gains over graph-regularization removal and graph-propagation removal are positive in every seed on Grocery, Reddit-S, Toys, Books, and Sports. The negative graph-regularization difference on Cloth also occurs in every seed.

Table 10: Complete single-modality semantic scores for the same controls. R@1 uses category relevance.

Variant	Grocery			Reddit-S		
	mAP@10	P@10	Sem. R@1	mAP@10	P@10	Sem. R@1
GOMA	**42.82** ± 0.22	**50.23** ± 0.17	60.16 ± 0.39	**59.74** ± 0.26	**65.39** ± 0.31	**70.13** ± 0.08
w/o graph regularization	40.69 ± 0.27	48.64 ± 0.15	**61.88** ± 0.38	52.94 ± 0.16	59.89 ± 0.12	68.89 ± 0.08
w/o graph propagation	42.41 ± 0.41	49.89 ± 0.33	60.09 ± 0.17	59.61 ± 0.17	65.28 ± 0.21	69.99 ± 0.19
Shared readout	39.68 ± 0.16	47.66 ± 0.11	60.87 ± 0.17	52.26 ± 0.45	59.36 ± 0.40	67.28 ± 0.29
w/o semantic mapping	39.53 ± 0.29	47.57 ± 0.27	60.63 ± 0.37	52.71 ± 0.47	59.78 ± 0.38	68.00 ± 0.17
Relabeled training graph	24.00 ± 1.19	32.88 ± 1.17	47.18 ± 1.44	53.28 ± 0.41	60.14 ± 0.34	67.40 ± 0.45
w/o content initialization	**42.82** ± 0.22	**50.23** ± 0.17	60.16 ± 0.39	59.65 ± 0.17	65.34 ± 0.26	70.06 ± 0.07
w/o semantic initialization	42.41 ± 0.41	49.89 ± 0.33	60.09 ± 0.17	59.42 ± 0.40	65.06 ± 0.39	70.04 ± 0.10
Content-guided edges	**42.82** ± 0.22	**50.23** ± 0.17	60.16 ± 0.39	59.61 ± 0.17	65.28 ± 0.21	69.99 ± 0.19

Variant	Toys			Books		
	mAP@10	P@10	Sem. R@1	mAP@10	P@10	Sem. R@1
GOMA	41.52 ± 0.44	50.22 ± 0.48	61.36 ± 0.30	46.16 ± 0.03	55.79 ± 0.01	59.79 ± 0.07
w/o graph regularization	36.32 ± 0.21	45.67 ± 0.19	59.61 ± 0.18	42.26 ± 0.02	52.22 ± 0.01	57.24 ± 0.11
w/o graph propagation	**41.85** ± 0.51	**50.60** ± 0.53	**61.44** ± 0.23	46.13 ± 0.09	55.78 ± 0.04	59.82 ± 0.14
Shared readout	37.36 ± 0.60	46.58 ± 0.62	59.96 ± 0.28	43.11 ± 0.10	53.23 ± 0.10	58.03 ± 0.09
w/o semantic mapping	37.43 ± 0.14	46.71 ± 0.13	60.18 ± 0.11	43.05 ± 0.14	53.20 ± 0.13	57.92 ± 0.05
Relabeled training graph	31.06 ± 4.24	40.25 ± 3.89	55.52 ± 3.30	43.81 ± 0.06	53.56 ± 0.06	58.10 ± 0.28
w/o content initialization	41.56 ± 0.39	50.29 ± 0.38	61.43 ± 0.25	46.08 ± 0.04	55.75 ± 0.03	59.73 ± 0.10
w/o semantic initialization	41.26 ± 0.16	49.99 ± 0.16	61.23 ± 0.10	46.11 ± 0.08	55.78 ± 0.01	59.75 ± 0.14
Content-guided edges	41.38 ± 0.22	50.08 ± 0.26	61.14 ± 0.11	**46.18** ± 0.05	**55.80** ± 0.02	**59.87** ± 0.07

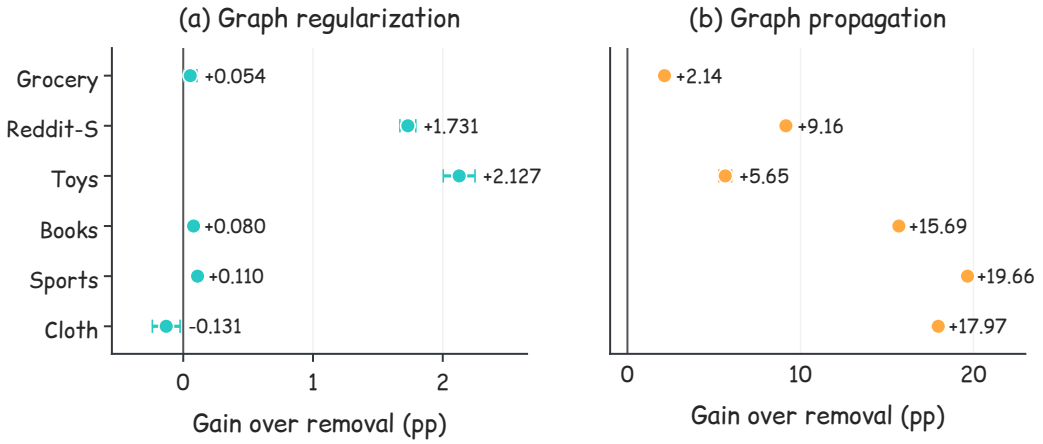

Figure 6: Paired GOMA gains over removal of graph regularization (a) and graph propagation (b). Points show means over seeds 43–45, and horizontal bars show the sample SD of paired seed differences. The first four datasets use semantic mAP@10 and the last two use relation MRR. Gains are in percentage points.

B.3 Propagation Interventions

Table 12 retains the selected GOMA parameters and changes only the indicated propagation computation. Removing propagation, removing restart, or relabeling the observed graph tests the use of observed relationships and the initial signal. Uniform edge weights retain the graph and initialization. Content guidance substitutes $\mathbf{C}$ for the semantic guide, and permuted semantic guidance changes only the guide-to-object assignment. The scale is estimated from the legal training edges before the diagnostic permutation. Initialization removals retain the full semantic-guided operator.

Table 11: Complete dual-attribute semantic and relation scores for the retrained controls.

Variant	Grocery			Reddit-S		
	mAP@10	P@10	Sem. R@1	mAP@10	P@10	Sem. R@1
GOMA	**49.60** ± 0.05	56.13 ± 0.03	67.82 ± 0.15	78.97 ± 0.05	81.80 ± 0.03	85.26 ± 0.14
w/o graph regularization	49.55 ± 0.09	**56.50** ± 0.11	**68.15** ± 0.32	77.24 ± 0.10	80.40 ± 0.09	85.72 ± 0.21
w/o graph propagation	47.46 ± 0.25	54.91 ± 0.19	67.24 ± 0.42	69.81 ± 0.18	74.63 ± 0.13	81.39 ± 0.42
Shared readout	49.05 ± 0.05	55.82 ± 0.07	67.63 ± 0.20	77.98 ± 0.16	81.12 ± 0.11	85.49 ± 0.28
w/o semantic mapping	48.98 ± 0.06	55.69 ± 0.13	67.51 ± 0.04	78.14 ± 0.07	81.18 ± 0.06	85.57 ± 0.09
Relabeled training graph	46.62 ± 0.07	53.61 ± 0.12	66.32 ± 0.34	77.12 ± 0.17	80.17 ± 0.18	85.72 ± 0.43
w/o content initialization	49.34 ± 0.18	55.90 ± 0.15	66.84 ± 0.13	**79.40** ± 0.31	**82.16** ± 0.25	85.74 ± 0.37
w/o semantic initialization	48.97 ± 0.08	55.93 ± 0.07	67.60 ± 0.06	76.67 ± 0.35	79.94 ± 0.30	**85.77** ± 0.07
Content-guided edges	49.20 ± 0.08	55.82 ± 0.04	67.31 ± 0.14	78.91 ± 0.08	81.76 ± 0.06	85.25 ± 0.11

Variant	Toys			Books		
	mAP@10	P@10	Sem. R@1	mAP@10	P@10	Sem. R@1
GOMA	51.31 ± 0.07	58.87 ± 0.07	67.72 ± 0.18	**66.70** ± 0.01	**73.19** ± 0.01	**78.12** ± 0.03
w/o graph regularization	49.18 ± 0.07	56.83 ± 0.09	67.36 ± 0.16	66.62 ± 0.03	73.02 ± 0.02	77.99 ± 0.04
w/o graph propagation	45.65 ± 0.41	54.28 ± 0.44	65.78 ± 0.36	51.01 ± 0.14	60.19 ± 0.11	66.43 ± 0.14
Shared readout	50.08 ± 0.25	57.67 ± 0.26	67.28 ± 0.30	66.53 ± 0.02	73.03 ± 0.02	77.98 ± 0.08
w/o semantic mapping	50.22 ± 0.11	57.81 ± 0.12	67.03 ± 0.23	66.46 ± 0.04	72.99 ± 0.03	77.90 ± 0.06
Relabeled training graph	47.62 ± 1.81	55.23 ± 1.81	66.65 ± 0.64	66.07 ± 0.06	72.67 ± 0.05	77.70 ± 0.09
w/o content initialization	**51.63** ± 0.24	**59.08** ± 0.31	**68.06** ± 0.19	66.56 ± 0.07	73.13 ± 0.06	**78.12** ± 0.13
w/o semantic initialization	49.05 ± 0.14	56.74 ± 0.15	67.10 ± 0.16	66.57 ± 0.06	73.01 ± 0.07	78.03 ± 0.08
Content-guided edges	50.99 ± 0.11	58.58 ± 0.17	67.50 ± 0.04	66.62 ± 0.03	73.12 ± 0.03	78.04 ± 0.05

Variant	Sports			Cloth		
	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10
GOMA	87.89 ± 0.03	80.98 ± 0.04	98.68 ± 0.03	81.32 ± 0.11	72.75 ± 0.13	95.70 ± 0.10
w/o graph regularization	87.78 ± 0.02	80.78 ± 0.03	98.63 ± 0.02	**81.46** ± 0.00	**72.95** ± 0.01	95.77 ± 0.03
w/o graph propagation	68.24 ± 0.11	56.40 ± 0.13	89.78 ± 0.09	63.36 ± 0.15	51.32 ± 0.13	86.50 ± 0.17
Shared readout	87.88 ± 0.04	80.97 ± 0.06	98.66 ± 0.05	81.34 ± 0.07	72.75 ± 0.09	95.74 ± 0.04
w/o semantic mapping	87.89 ± 0.02	81.02 ± 0.04	98.64 ± 0.04	81.38 ± 0.07	72.81 ± 0.11	95.76 ± 0.08
Relabeled training graph	**88.00** ± 0.04	**81.12** ± 0.02	**98.73** ± 0.05	81.36 ± 0.01	72.76 ± 0.02	95.76 ± 0.07
w/o content initialization	87.88 ± 0.03	80.94 ± 0.02	98.68 ± 0.04	81.40 ± 0.02	72.83 ± 0.04	**95.79** ± 0.01
w/o semantic initialization	87.91 ± 0.02	81.00 ± 0.04	98.67 ± 0.03	81.37 ± 0.04	72.81 ± 0.06	95.75 ± 0.01
Content-guided edges	87.91 ± 0.03	81.02 ± 0.07	98.67 ± 0.03	81.39 ± 0.06	72.81 ± 0.08	95.77 ± 0.03

The observed graph improves all six primary measures over graph removal and graph relabeling. Semantic guidance improves mAP@10 over uniform weights, content guidance, and permuted semantic guidance on every category dataset, with positive differences in each seed. On Sports and Cloth, uniform and content-guided weights give higher MRR. Thus the additional benefit of the learned semantic guide is supported most directly by category retrieval. GOMA's relation results remain above the strongest external comparisons.

Removing restart reduces the primary measure on five datasets and slightly improves Sports MRR. Removing the content or semantic part of the initialization also gives mixed changes. The full initialization and restart preserve access to both learned outputs, while their ranking effect depends on the retrieval target and graph. The tables retain these differences alongside the signal measurements.

B.4 Empirical Probes

For graph-category agreement, we remove self edges and collapse repeated undirected training edges. The observed same-category proportion is deterministic. Three relabelings use seeds 1001, 1002, and 1003 and permute vertices within exact training-degree groups. Each object retains its degree in this descriptive control. The resulting SD measures graph-relabeling variation, separately from training-seed variation.

Table 12: Complete propagation interventions at the same GOMA weights. Only the named computation changes.

Propagation setting	Grocery			Reddit-S		
	mAP@10	P@10	Sem. R@1	mAP@10	P@10	Sem. R@1
GOMA	49.60 ± 0.05	56.13 ± 0.03	67.82 ± 0.15	78.97 ± 0.05	81.80 ± 0.03	85.26 ± 0.14
w/o graph propagation	47.59 ± 0.08	54.96 ± 0.07	67.63 ± 0.92	69.83 ± 0.16	74.66 ± 0.12	81.23 ± 0.44
w/o restart	43.31 ± 0.13	50.13 ± 0.11	58.65 ± 0.19	77.59 ± 0.06	80.68 ± 0.04	84.35 ± 0.21
Relabeled observed graph	36.32 ± 0.61	45.04 ± 0.63	56.18 ± 0.54	29.08 ± 0.64	40.79 ± 0.55	40.66 ± 0.55
Uniform edge weights	48.85 ± 0.09	55.50 ± 0.02	66.86 ± 0.23	78.95 ± 0.08	81.79 ± 0.06	85.36 ± 0.13
Content-guided edges	49.20 ± 0.08	55.82 ± 0.04	67.31 ± 0.14	78.93 ± 0.07	81.79 ± 0.05	85.26 ± 0.11
Permuted semantic guide	48.89 ± 0.11	55.59 ± 0.08	66.78 ± 0.11	78.95 ± 0.06	81.81 ± 0.04	85.31 ± 0.19
w/o content initialization	49.34 ± 0.18	55.90 ± 0.15	66.84 ± 0.13	79.28 ± 0.13	82.06 ± 0.10	85.70 ± 0.31
w/o semantic initialization	48.92 ± 0.08	55.94 ± 0.07	67.31 ± 0.24	76.44 ± 0.28	79.77 ± 0.25	85.64 ± 0.19
Propagation setting	Toys			Books		
	mAP@10	P@10	Sem. R@1	mAP@10	P@10	Sem. R@1
GOMA	51.31 ± 0.07	58.87 ± 0.07	67.72 ± 0.18	66.70 ± 0.01	73.19 ± 0.01	78.12 ± 0.03
w/o graph propagation	45.43 ± 0.21	54.10 ± 0.20	65.58 ± 0.26	51.09 ± 0.13	60.26 ± 0.08	66.48 ± 0.03
w/o restart	49.50 ± 0.11	57.10 ± 0.09	65.04 ± 0.11	65.14 ± 0.03	71.63 ± 0.03	76.44 ± 0.05
Relabeled observed graph	31.24 ± 0.38	41.67 ± 0.45	50.27 ± 1.10	47.13 ± 0.29	56.84 ± 0.18	63.39 ± 0.36
Uniform edge weights	50.95 ± 0.10	58.50 ± 0.10	67.18 ± 0.37	66.52 ± 0.03	73.05 ± 0.04	77.97 ± 0.09
Content-guided edges	51.07 ± 0.11	58.63 ± 0.12	67.62 ± 0.23	66.63 ± 0.02	73.13 ± 0.02	77.99 ± 0.04
Permuted semantic guide	50.90 ± 0.09	58.48 ± 0.06	67.14 ± 0.22	66.53 ± 0.03	73.06 ± 0.02	77.92 ± 0.06
w/o content initialization	51.64 ± 0.21	59.10 ± 0.26	68.11 ± 0.10	66.56 ± 0.05	73.12 ± 0.04	78.08 ± 0.12
w/o semantic initialization	48.92 ± 0.22	56.64 ± 0.19	67.19 ± 0.14	66.55 ± 0.06	72.98 ± 0.06	78.10 ± 0.05
Propagation setting	Sports			Cloth		
	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10
GOMA	87.89 ± 0.03	80.98 ± 0.04	98.68 ± 0.03	81.32 ± 0.11	72.75 ± 0.13	95.70 ± 0.10
w/o graph propagation	67.93 ± 0.11	55.99 ± 0.13	89.66 ± 0.04	63.36 ± 0.13	51.12 ± 0.32	86.84 ± 0.16
w/o restart	88.32 ± 0.01	81.63 ± 0.00	98.80 ± 0.05	81.27 ± 0.10	72.98 ± 0.12	95.42 ± 0.09
Relabeled observed graph	50.63 ± 2.93	38.24 ± 2.73	74.45 ± 3.65	48.37 ± 2.70	36.69 ± 2.25	71.52 ± 3.44
Uniform edge weights	88.40 ± 0.02	81.59 ± 0.04	98.91 ± 0.02	81.96 ± 0.10	73.48 ± 0.15	96.06 ± 0.06
Content-guided edges	87.91 ± 0.03	81.00 ± 0.05	98.69 ± 0.03	81.35 ± 0.09	72.77 ± 0.10	95.73 ± 0.07
Permuted semantic guide	88.37 ± 0.02	81.54 ± 0.04	98.89 ± 0.03	81.89 ± 0.09	73.35 ± 0.12	96.05 ± 0.07
w/o content initialization	87.89 ± 0.04	81.00 ± 0.10	98.68 ± 0.04	81.28 ± 0.14	72.71 ± 0.14	95.65 ± 0.14
w/o semantic initialization	87.91 ± 0.03	80.99 ± 0.04	98.70 ± 0.01	81.37 ± 0.08	72.81 ± 0.11	95.74 ± 0.07

The shared-output probe uses two residual mappings per modality. Identity ranking, paired cross-entropy, and an unnormalized edge-average graph regularizer supervise its final representation. Intermediate and final outputs are normalized and scored using the same validation candidates at the jointly selected model. This is a representation probe of shared supervision, separate from the final GOMA comparison. Table 13 reports final-minus-intermediate changes and the current GOMA embedding measurements. Table 4 tests the distinct question of whether supervision at separate outputs helps when parameter counts are equal.

The propagation probe compares graph regularization with its removal at a common training step per dataset and seed. Both variants initialize from $\mathcal{N}(\mathbf{C} + \mathbf{Z})$ and use the same eight-step restart rule with uniform observed-edge weights. The common step is the earlier of their selected training steps, as specified before this probe. We evaluate the initial and propagated representations using the same candidates. Table 14 gives the graph-regularization gain before and after propagation and their difference. Every initial gain is positive, every interaction is negative, and propagation improves both variants in all seeds. This identifies overlapping benefits without assigning a unique mechanism to the negative interaction.

Fig. 4(a) averages raw squared endpoint differences over feature coordinates and modalities, then computes each seed's semantic-embedding reduction relative to its content embedding. This diagnostic differs from the degree-normalized training regularizer. Identity curves first average the two

Table 13: Graph-category agreement, shared-output probe changes, and GOMA validation embeddings. Graph agreement uses three relabelings and retrieval uses three model seeds.

Quantity	**Grocery**	**Reddit-S**	**Toys**	**Books**
Observed same-category edges (%)	68.21	96.32	72.68	67.78
Relabeled same-category edges (%)	9.66 ± 0.09	73.65 ± 0.05	9.71 ± 0.18	16.20 ± 0.01
Probe final minus intermediate R@1 (Validation)	-0.99 ± 0.17	-1.10 ± 0.30	-1.93 ± 0.17	0.84 ± 0.59
Probe final minus intermediate mAP@10	1.18 ± 0.17	1.42 ± 0.22	2.73 ± 0.07	0.82 ± 0.11
Content embedding identity R@1	52.79 ± 0.19	18.62 ± 0.02	63.57 ± 0.26	50.65 ± 0.57
Content embedding semantic mAP@10	33.12 ± 0.42	39.55 ± 1.30	30.70 ± 0.80	37.28 ± 0.39
Semantic embedding identity R@1	39.79 ± 0.18	13.07 ± 0.19	48.55 ± 0.94	38.05 ± 1.89
Semantic embedding semantic mAP@10	43.48 ± 0.36	58.63 ± 0.20	41.41 ± 0.41	43.33 ± 0.14

Table 14: Graph-regularization gains before and after uniform graph propagation at matched training steps. The first four columns use mAP@10 and the last two use MRR. Entries report paired means and sample SDs in percentage points.

Graph-regularization gain	**Grocery**	**Reddit-S**	**Toys**	**Books**	**Sports**	**Cloth**
Before propagation	2.61 ± 0.28	5.29 ± 0.35	6.37 ± 0.20	2.52 ± 0.12	0.48 ± 0.08	0.31 ± 0.11
After uniform propagation	-0.39 ± 0.02	1.44 ± 0.23	1.69 ± 0.08	-0.11 ± 0.04	0.09 ± 0.01	-0.06 ± 0.05
After minus before	-3.01 ± 0.28	-3.85 ± 0.20	-4.68 ± 0.26	-2.63 ± 0.15	-0.39 ± 0.09	-0.36 ± 0.16

directional empirical rank distributions within each dataset, then give the four category datasets equal weight. Their mean and band summarize three seed curves. The edge-weight curves in Fig. 5(a) average empirical cumulative distributions over seeds within each dataset. Both energy measurements in Fig. 5(c) use that model's semantic-guided operator.

C Mathematical Analysis

C.1 Normalized Graph Dirichlet Energy

Let $\mathbf{A}_{\mathcal{T}}$ be the binary symmetric training adjacency and $\mathbf{D}_{\mathcal{T}}$ its degree matrix. On non-isolated training nodes, define the normalized Laplacian $\mathbf{L}_{\mathcal{T}} = \mathbf{I} - \mathbf{D}_{\mathcal{T}}^{-1/2}\mathbf{A}_{\mathcal{T}}\mathbf{D}_{\mathcal{T}}^{-1/2}$. Isolated nodes contribute zero to the edge energy. The normalized Dirichlet identity is

$$\operatorname{tr}(\mathbf{Z}^\top \mathbf{L}_{\mathcal{T}} \mathbf{Z}) = \sum_{\{i,j\}\in\mathcal{E}_{\mathcal{T}}} \left\| \frac{\mathbf{z}_i}{\sqrt{d_i}} - \frac{\mathbf{z}_j}{\sqrt{d_j}} \right\|_2^2. \tag{7}$$

The sum here contains each undirected edge once. Expanding it gives degree-normalized squared norms and adjacency cross terms. Equation 2 averages the two modality energies over feature coordinates and edges. A bidirected edge list repeats each term twice in both the numerator and denominator, giving the same average. Degrees come from the complete training-induced graph, independently of the sampled edge batch.

Under the unit-norm representation used by GOMA, each edge term can also be written as

$$\left\| \frac{\mathbf{z}_i}{\sqrt{d_i}} - \frac{\mathbf{z}_j}{\sqrt{d_j}} \right\|_2^2 = \frac{1}{d_i} + \frac{1}{d_j} - \frac{2\,\mathbf{z}_i^\top \mathbf{z}_j}{\sqrt{d_i d_j}}. \tag{8}$$

Thus the energy rewards degree-weighted cosine alignment together with degree-dependent norm terms. Zero energy means that $\mathbf{z}_i/\sqrt{d_i}$ is constant on each connected component. On an irregular degree graph, this is not the same as a simple constant embedding over the whole graph, and unit-norm embeddings cannot attain it across an edge whose endpoint degrees differ.

For $\mathbf{L}_{\mathcal{T}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, this energy equals $\sum_\ell \lambda_\ell \|\mathbf{u}_\ell^\top \mathbf{Z}\|_2^2$. It penalizes variation associated with larger eigenvalues of the training Laplacian. The content embedding is a jointly learned representation with direct identity supervision. It is not constructed as a complementary frequency band. Both its mapping and the semantic mapping receive gradients from the semantic objective.

C.2 Restart Propagation

Write $\beta = 1 - \alpha$. Expanding Equation 5 by induction gives

$$\mathbf{H}_K = \left[\alpha \sum_{j=0}^{K-1} \beta^j \mathbf{P}_W^j + \beta^K \mathbf{P}_W^K \right] \mathbf{H}_0. \tag{9}$$

For $K = 1$, the expression is $\alpha\mathbf{H}_0 + \beta\mathbf{P}_W\mathbf{H}_0$, which is the first update. Substituting the expression for K into the next update adds the $j = 0$ term and shifts each existing power by one. This establishes the formula for all positive integer depths. The edge weights have already been computed from the learned semantic signal and remain unchanged during these iterations. The result is a polynomial in that operator, with no eigendecomposition required during execution.

Because semantic edge weights are positive and symmetric, the weighted normalized adjacency on non-isolated components has spectrum in $[-1, 1]$. Adding the identity and dividing by two places the spectrum of $\mathbf{P}_W$ in $[0, 1]$. Isolated nodes use eigenvalue one by construction. For $0 < \alpha \leq 1$, the polynomial p_K has nonnegative coefficients, is nondecreasing on $[0, 1]$, satisfies $p_K(1) = 1$, and lies between α and one for $K \geq 1$. Hence the finite filter emphasizes components with smaller eigenvalues of $\mathbf{I} - \mathbf{P}_W$, while maintaining a positive contribution from the initial signal in every frequency component.

As depth increases, the geometric series converges because $\beta < 1$. The limit is

$$\mathbf{H}_\infty = \alpha(\mathbf{I} - \beta\mathbf{P}_W)^{-1}\mathbf{H}_0. \tag{10}$$

It is also the unique minimizer of

$$\alpha\|\mathbf{H} - \mathbf{H}_0\|_F^2 + \beta\,\mathrm{tr}\big(\mathbf{H}^\top(\mathbf{I} - \mathbf{P}_W)\mathbf{H}\big)\,. \tag{11}$$

Differentiation yields $(\alpha\mathbf{I} + \beta(\mathbf{I} - \mathbf{P}_W))\mathbf{H} = \alpha\mathbf{H}_0$, which is equivalent to Equation 10. Positive α makes the quadratic strictly convex. GOMA uses the eight-step finite filter in the reported experiments. These identities explain restart propagation for a given weighted operator (Klicpera et al., 2019). The operator itself depends nonlinearly on the learned semantic signal, so this conditional quadratic does not describe the entire learned model.

C.3 Normalization and Graph Operators

The final propagated representation is $\widehat{\mathbf{H}} = \mathcal{N}(\mathbf{H}_K)$. Row normalization scales nodes by factors determined by their filtered norms. It generally does not commute with $\mathbf{P}_W$, so the final representation is not described solely by the polynomial response in Equation 9. Fig. 5(c) reports the measured energy on both sides of this operation. With the relative energy defined by $\mathrm{tr}(\mathbf{H}^\top(\mathbf{I}-\mathbf{P}_W)\mathbf{H})/\|\mathbf{H}\|_F^2$, changes in the figure are properties of the actual signals used in the computation.

Because every row of $\widehat{\mathbf{H}}$ has unit norm, its final inner product has a comparable cosine-style scale across objects. Row normalization is an independent nonlinear step applied after propagation. The low-pass polynomial in Equation 9 describes only the unnormalized signal $\mathbf{H}_K$, so it cannot directly describe the final normalized representation.

Graph regularization and graph propagation use different forms of graph smoothness. Training minimizes an edge-average energy through nonlinear, normalized semantic mappings on training objects. Propagation applies a semantic-guided lazy symmetric operator to the combined content and semantic initialization over the permitted observed graph. The first transfers relational information into independently computable attribute representations. The second combines these representations with observed neighborhoods when that information is available. The learned semantic signal connects them by contributing to both initialization and edge weights. The training-induced and observed retrieval graphs can contain different nodes and edges, and propagation uses weighted degrees.

Algorithm 1 GOMA training and retrieval

Require: Paired features, training objects $\mathcal{T}$, observed graph, validation pools
Ensure: One trained model and its task-appropriate retrieval embeddings
Form the induced training graph and the permitted observed graph
Initialize f_m, g_m with zero final linear layers for each modality m
Set η, λ from gradient norms on ten initialization batches
for each optimization update **do**
    Sample paired training objects and training edges independently
    Compute $\mathbf{c}^m, \mathbf{z}^m$ for sampled objects and edge endpoints
    Evaluate Equation 3 and update f_m, g_m jointly with AdamW
    if the update is a positive multiple of five **then**
        Evaluate the joint validation score or relation MRR
        Save the model on a strict improvement
    end if
end for
Select one model using the dataset's validation rule
Single-modality retrieval
Encode each query and candidate using its own supplied attribute
Use content inner products for identity or semantic inner products for cross-object relevance
Dual-attribute retrieval
Fuse both attributes into $\mathbf{C}$ and $\mathbf{Z}$, then form $\mathbf{H}_0 = \mathcal{N}(\mathbf{C} + \mathbf{Z})$
Compute the semantic edge scale on legal training edges
Weight observed edges using Equation 4 and form $\mathbf{P}_W$
Apply eight propagation steps using Equation 5 and normalize the result
Cache the final representations in the offline retrieval index and rank by inner products

D Algorithm and Implementation

D.1 Objectives and Loss Weights

GOMA directly supervises the content embedding with hard-negative identity ranking (Faghri et al., 2017). For a batch of B paired objects, define $s^c_{ij} = (\mathbf{c}^v_i)^\top \mathbf{c}^t_j$. The identity objective is

$$\mathcal{L}_{\text{id}} = \frac{1}{B} \sum_{i=1}^{B} \left(\left[\gamma + \max_{j \neq i} s^c_{ij} - s^c_{ii} \right]_+ + \left[\gamma + \max_{j \neq i} s^c_{ji} - s^c_{ii} \right]_+ \right), \tag{12}$$

where $[a]_+ = \max(a, 0)$ and $\gamma = 0.2$. Each positive is compared with its most competitive incorrect candidate in both directions. This directly supervises identity at the earlier embedding. The semantic loss described next also updates f_m through g_m, so both embeddings share learned content and are optimized together.

Algorithm 1 summarizes the computation. Each f_m and g_m contains two d-wide linear layers with a GELU between them. Both output layers are initialized to zero, so the residual paths initially reproduce the normalized input. The weights are independent across modalities and stages. All four mappings are jointly optimized, and there is no stop-gradient operation between the content and semantic embeddings.

For semantic scores $s^z_{ij} = (\mathbf{z}^v_i)^\top \mathbf{z}^t_j$ in a batch of B paired objects, the paired objective used in Equation 3 is

$$\mathcal{L}_{\text{pair}} = -\frac{1}{2B} \sum_{i=1}^{B} \left(\log \frac{\exp(s^z_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s^z_{ij}/\tau)} + \log \frac{\exp(s^z_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s^z_{ji}/\tau)} \right). \tag{13}$$

The two terms give equal weight to image-to-text and text-to-image matching. Identity ranking uses Equation 12 with margin 0.2. The implementation averages the hard-negative loss over a batch while summing its two directions. The graph regularizer averages sampled degree-normalized endpoint differences over the two modalities and feature coordinates. It therefore requires no dense query-edge score matrix.

We set the loss weights from ten training-only batches at initialization. Let a_b be the norm of the identity-loss gradient with respect to the content mappings, b_b the norm of the paired-loss gradient with respect to the semantic mappings, and c_b the norm of the graph-energy gradient with respect to

the semantic mappings. If bars denote means over those ten batches, the coefficients are $\eta = \bar{a}/\bar{b}$ and $\lambda = \bar{b}/\bar{c}$. Finite nonzero means are required for a nonempty training graph. The coefficients are stored and used throughout training, including the active terms of the component controls. This sets the initial gradient scales without introducing an additional learned weighting mechanism.

D.2 Sampling and Model Selection

Each update samples up to 1,024 paired training objects without replacement and 1,024 training edges with replacement. The union of these object IDs is encoded once per modality. The training edge list is restricted to pairs of training endpoints before sampling. AdamW uses learning rate 0.001 and weight decay 0.0001 (Loshchilov & Hutter, 2019). The maximum trajectory length is 500 updates, and validation is performed every five updates. The first eligible selected model is at update five, and equal validation values retain the earlier model.

For the semantic datasets, the selected score is the arithmetic mean of bidirectional identity R@1, single-modality semantic mAP@10, and dual-attribute semantic mAP@10. This gives the three applicable task objectives one common selection rule. Relation datasets use MRR. Patience ends a selection trajectory after 50 optimization updates without strict improvement. Additional task-specific validation states were recorded for analysis, but every GOMA number in the comparison tables uses only the common selected model. Training can continue while other tracked validation objectives remain active, within the same 500-update maximum.

D.3 Computational Cost

Table 15 reports the number of trainable parameters and the observed training cost for the GOMA runs. The parameter count excludes the frozen upstream feature encoders. The update count covers the recorded training trajectory, including updates after the selected iteration. Runtime measures the recorded training and validation trajectory under the shared GPU execution environment. Peak memory is the maximum allocation reported by PyTorch during the run. These measurements describe the execution cost of the reported implementation. Fig. 7 separately compares the stages of complete-record index construction after a selected checkpoint. It uses one warmup and three synchronized repeats for each of seeds 43–45 on the recorded NVIDIA RTX PRO 6000, excludes checkpoint and disk loading, and does not run training, validation, or report scoring. The stage and whole-index timers are measured separately, so their means are not additive.

Table 15: GOMA training cost over three seeds. The shared GPU environment makes runtime dependent on concurrent workload.

Dataset	Parameters	Updates	Training seconds	Peak allocated GiB
Grocery	2,101,248	243.3 ± 5.8	9.89 ± 0.14	0.75 ± 0.02
Reddit-S	2,101,248	235.0 ± 25.0	9.34 ± 0.90	0.72 ± 0.02
Toys	2,101,248	271.7 ± 80.2	12.66 ± 3.26	0.84 ± 0.03
Books	132,096	435.0 ± 58.9	35.55 ± 4.70	5.91 ± 0.00
Sports	132,096	56.7 ± 2.9	0.70 ± 0.03	0.83 ± 0.68
Cloth	132,096	60.0 ± 5.0	0.93 ± 0.08	1.06 ± 0.00

The four residual mappings contain $8d^2 + 8d$ trainable parameters. They require $O(nd^2)$ arithmetic when encoding n objects, with a constant number of layers. A paired batch additionally forms B^2 similarities, while the graph regularizer uses sampled endpoint differences. The object index stores $O(Nd)$ representations and $O(N + |\mathcal{E}|)$ sparse graph entries. Computing edge weights takes one pass over semantic endpoint distances. Each propagation round is a sparse matrix multiplication followed by a weighted addition of the initial signal. Query scoring can be performed in chunks without changing the candidate list or relevance evaluation. The implementation uses PyTorch and NumPy, with one common random seed controlling model initialization and training samples.

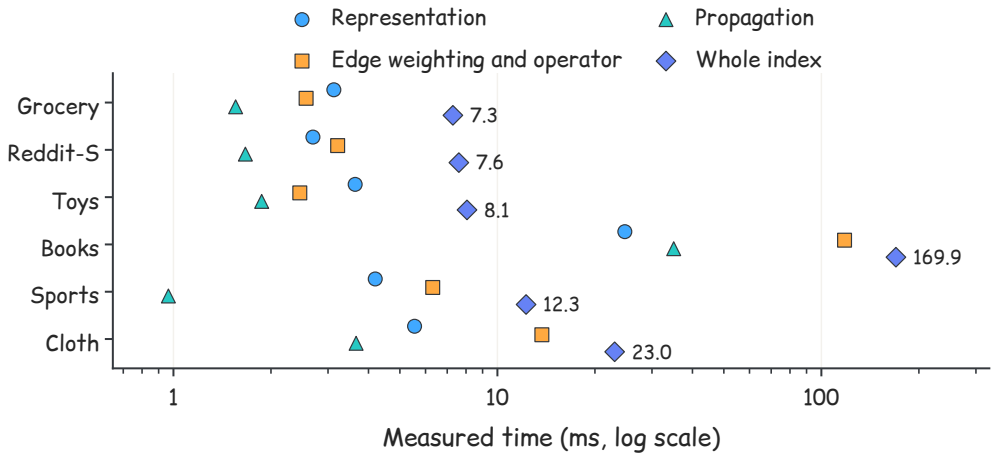

Figure 7: Offline index construction time for selected GOMA checkpoints. Points show means over seeds 43–45 after one warmup and three synchronized repeats per seed. Stage and whole-index times were measured separately and need not sum. I/O and checkpoint loading are excluded.

E Data, Splits, and Evaluation

E.1 Features and Observed Graphs

The experiments use multimodal graph data and released feature resources associated with MAGB and OpenMAG (Yan et al., 2024; Wan et al., 2026). Table 16 summarizes the object and evaluation sizes. Grocery and Toys use item relationships from the `also_buy` and `also_view` metadata. Reddit-S uses the released co-comment relationship graph. These relationships are supplied independently of the category labels used for semantic evaluation. The graph operators use symmetrized binary edges with duplicate edges and self edges removed.

Table 16: Dataset, observed-graph, and retrieval-pool sizes. Edge entries count both orientations. Semantic galleries have the same size as their query pools. Relation queries each rank one positive among 150 negatives.

Quantity	**Grocery**	**Reddit-S**	**Toys**	**Books**	**Sports**	**Cloth**
Objects	17,074	15,894	20,695	685,294	50,250	125,839
Observed edge entries	142,262	283,080	113,402	13,682,358	603,696	1,622,678
Feature width	512	512	512	128	128	128
Training objects	10,244	9,536	12,417	406,689	30,150	75,503
Validation queries	1,702	1,584	2,064	4,000	4,096	4,096
Evaluation queries	1,702	1,584	2,064	10,000	9,825	9,917

Grocery, Reddit-S, and Toys use frozen 512-dimensional image and text projections of Qwen2-VL features (Wang et al., 2024). Their upstream paired content models are trained without graph propagation, and all compared methods use the same projected arrays. Sports and Cloth use released 512-dimensional CLIP feature caches projected to 128 dimensions with upstream seed 44. The released cache metadata does not specify the exact encoder-weight revision. Books uses Qwen2-VL features projected to 128 dimensions by a paired content model with upstream seed 44. GOMA learns its content and semantic mappings on these common prepared features.

E.2 Semantic Retrieval

Grocery, Reddit-S, and Toys use common training, validation, and report identity partitions. Within each validation or report partition, the objects of each category are shuffled and divided into equal query and gallery halves. At most one object in an odd-sized category is unused, and no category

is removed by a minimum-size threshold. This produces disjoint query and gallery objects while retaining all represented categories.

Books retains the released training object IDs and the ten official assigned labels 0–9 provided by the release. For each class, validation uses 400 query and 400 gallery objects, while test reporting uses 1,000 query and 1,000 gallery objects. The sampling seeds are 4401 and 4402, respectively. The upstream paired projection is selected before the compared feature-adaptation methods are trained. Semantic retrieval supplies the observed graph and evaluates category relevance on the reporting objects.

Semantic retrieval averages image-to-text and text-to-image evaluation in both orientations of the disjoint query/gallery pools. For dual-attribute representations, the two modality labels yield the same score, while the two query/gallery orientations remain distinct. A candidate is relevant when it has the same category as the query and a different object identity. For a query with R relevant candidates, AP@10 divides the sum of precision at relevant ranks up to ten by $\min(10, R)$. Precision at ten is the relevant fraction of the first ten candidates, and semantic R@1 records whether the highest-ranked candidate is relevant. Scores are first averaged within categories and then across categories.

E.3 Identity Retrieval

Identity retrieval takes the union of the corresponding semantic query and gallery IDs. Every image queries all text attributes in this union, and every text queries the image attributes. The matching object ID defines the true pair. Candidate ordering is permuted using seed 918273. IDs define relevance and candidate ordering but are not input features to GOMA. Each identity query has equal weight, and the reported primary recall averages the two modality directions. Thus an identity pool contains twice the number of objects in one semantic query half.

The strict single-modality interface also applies to candidates. An image representation is computed from image features alone, and a text representation from text features alone. In particular, the presence of a paired attribute elsewhere in the data store does not make that attribute part of the query encoder's input. Training can still use known pairs from the training partition, as in ordinary cross-modal representation learning.

E.4 Relation Ranking

Sports and Cloth use canonical unordered relation candidates with pool-construction seed 20260905 and a seeded random orientation for each positive. Validation contains 4,096 positive queries. Report construction begins with 10,000 positives and removes overlaps with the complete validation candidate pool, leaving 9,825 Sports queries and 9,917 Cloth queries. Target edges and reverse orientations are removed from the observed adjacency before propagation.

Each positive is ranked against 150 distinct negatives. Negative endpoints match the positive endpoint's degree bin, defined by $\lfloor \log_2(d+1) \rfloor$, expanding the bin radius when necessary. Self candidates and known positive relationships are excluded. Additional graph-support realizations are used only to exclude known positives in candidate construction and are not additional model inputs. All methods rank the same positive and negative arrays.

For positive score s^+ and negative scores s_j^-, the tie-aware rank is $r = 1 + \frac{1}{2}(\sum_j \mathbb{1}[s_j^- > s^+] + \sum_j \mathbb{1}[s_j^- \geq s^+])$. MRR averages $1/r$, while Hits@k averages $\mathbb{1}[r \leq k]$. A queried object and its true endpoint can retain their other observed relationships. This is the relational context available to the dual-attribute model when predicting the missing link.

F External Methods and Their Retrieval Adaptations

F.1 Common Inputs and Objectives

All external numbers are obtained from runs using the same prepared features and retrieval pools as GOMA. The roster is identical for single-modality and dual-attribute retrieval. The comparison evaluates adaptations of representation, recommendation, clustering, and multimodal graph models

through their output embeddings. Each method keeps its formal name, and the following details specify how its original input and output computations are used for these retrieval tasks.

VSE++ uses independent linear maps and margin-0.2 max-violation ranking. Its implementation averages the batch loss while summing the two directional terms. CLIP-Refine uses identity-initialized linear maps and its feature-level refinement objectives. The weights are zero for its additional contrastive term and one for each random-classification and knowledge-distillation term, with its $\alpha = 0.5$ and $T = 1$ settings. SigLIP uses independent modality maps with its pairwise sigmoid objective and learned logit scale and bias. These three methods are direct vision-language comparators for single-modality identity retrieval. GCN, GAT, GraphSAGE, and the multimodal graph methods instead serve as missing-modality adaptation controls. They retain their graph structure and auxiliary objectives while using independently computable views when one attribute is unavailable.

F.2 Dual-Attribute Retrieval

GCN, GAT, and GraphSAGE operate on concatenated image and text features and provide a shared node representation. MMGCN and MGAT retain modality-specific graph computation, while LGMRec retains its local and global graph construction. GCN, GAT, GraphSAGE, MMGCN, MGAT, LGMRec, CAMPA, and RoleMAG use adapted paired supervision for retrieval. RoleMAG retains its role-expert computation with the released zero auxiliary-loss coefficient. Their permitted fused or node outputs are used for dual-attribute inner-product ranking.

DMGC retains its reconstruction and clustering objectives and uses the shared representation before the classifier for scoring. The cluster count uses the known category count on semantic datasets and the inherited two-cluster setting on relation datasets. DGF retains feature filtering and graph/community objectives, with paired supervision restricted to training objects. CoMAG retains its multihop computation and shared/private representation choices under validation selection. These method-specific objectives remain active in their retrieval adaptations.

Dual-attribute external runs use a maximum of 500 updates and validation every five updates. Optimization uses Adam with nominal learning rate 0.001 and weight decay 0.0001. Pair batches can contain up to 16,384 objects. Most runs use 50-update patience, while Sports DMGC, DGF, and CoMAG use 25. Books generally uses 25, with DGF using 50. We retain each method's established training-selection protocol, while using a common maximum of 500 updates and validation every five updates. CoMAG's width-512, depth-four configuration is selected from four validation configurations on Sports. The same retained external runs supply every dual-attribute comparison in this paper.

F.3 Single-Modality Retrieval

Single-modality encoding supplies only the visible attribute of one query or candidate object. Training retains each graph method's layers, modality branches, aggregation or filtering operations, and auxiliary objectives. Paired retrieval supervision acts on their training graph outputs alongside each method's own auxiliary objectives. At inference, each independently encoded object uses the method's corresponding missing-input computation.

GCN, GAT, and GraphSAGE retain their graph layers and output heads, with the missing attribute set to zero in the single-object view. MMGCN and MGAT use their implementations' no-ID option. LGMRec keeps the local training graph and global hypergraph during learning, while a query uses its single-object hypergraph computation. DMGC keeps its shared low-frequency graph, modality-specific high-frequency processing, fusion, reconstruction, and clustering objectives during training. For a query, the low-frequency computation includes the object itself and the empty-neighborhood high-frequency computation corresponds to an identity Laplacian. DGF directly uses its independent modality output together with the graph and paired objectives during training.

CAMPA retains propagation and self/cross-trajectory attention. A missing trajectory starts from zero, and an unavailable neighborhood uses a zero propagation operator. CoMAG retains its full multihop matcher, shared/private representations, and orthogonality objective. Its single-modality query invokes the released model's graph-free path with the missing attribute set to zero, so query representations are computed independently. RoleMAG uses its missing-modality mask and role-based computation.

These adaptations provide independent query representations without constructing a graph between the evaluation queries or accessing their missing attributes.

For the 11 graph-based adaptations, each training subgraph contains the sampled paired objects and sampled edge endpoints, together with all training edges induced by their union. Batches use 1,024 pairs and 1,024 sampled edges. Training uses Adam with learning rate 0.001, weight decay 0.0001, and at most 500 updates. DMGC and DGF retain coefficient one for their auxiliary objectives. CoMAG uses its own retrieval loss with temperature 0.1 and its configured orthogonality coefficient 0.01. Identity R@1 and semantic mAP@10 have separate validation selectors with 50-update patience and five-update evaluation intervals. The first eligible model is at update five, and ties retain the earliest model. VSE++, CLIP-Refine, and SigLIP use the same single-modality pair-batch size and trained-model selection rule. The task-specific external selections are retained in the comparison with GOMA's one jointly selected model.

G LIMITATIONS

GOMA leads all 14 primary measures against the 14 external methods on the common prepared features. Complete directional and higher-rank recall results are reported in Table 5. All reported models are selected using the predefined validation protocol, while evaluation on additional unseen collections would further test cross-dataset generalization. Component behavior depends on the retrieval task. Semantic guidance is most consistently supported on category retrieval, while relation-ranking graphs can favor simpler edge weights. The deployment setting uses complete image-text records and a permitted observed graph to construct an offline index. Changes to the graph or learned representations require rebuilding that index. Online insertion and raw feature extraction lie outside this setting. The benchmark supplies the retrieval intent with each query rather than inferring it automatically.