OCG: ORTHOGONAL CONSENSUS GRAPH LEARNING FOR ROBUST HETEROGENEOUS REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Heterogeneous graph representation learning uses typed relations and node attributes, both of which can be corrupted. We propose Orthogonal Consensus Graph Learning (OCG), which augments contrastive alignment between matched nodes with a residual cross-Gram penalty and stochastic-path predictive consistency. We derive an exact decomposition of the residual penalty into a spectral shrinkage term for the semantic residual and a coupling term for its interaction with relational disagreement. We derive bounds on corruption-induced changes in the semantic residual and the fused representation under directional separation conditions. On ACM, DBLP, Yelp, and MAG, OCG achieves the highest mean Macro-F1 and Micro-F1 among the evaluated baselines. We evaluate robustness under edge and node-feature perturbations, assess individual objectives through ablation, and examine residual geometry in a controlled experiment.

1 INTRODUCTION

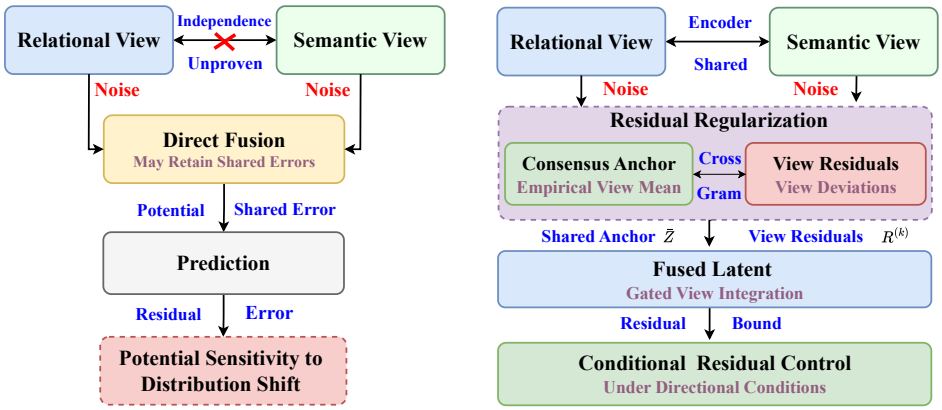
Heterogeneous graphs describe entities through typed relations and associated attributes, providing multiple sources of information for representation learning (Zhang et al., 2019, Sun et al., 2011, Dong et al., 2017). In an academic network, for example, papers can be related through shared authors or subjects, with each relation defining a different neighborhood for the same paper. Relation-induced views over a common node set form a multiplex representation (Kivelä et al., 2014, Boccaletti et al., 2014), allowing learners to combine information from several relational contexts for downstream prediction (Qian et al., 2024, Shen et al., 2024). Attribute-derived neighborhoods provide an additional view of feature-space proximity, complementing the connections specified by relational schemas.

Corruption in graph structure and node attributes can affect representations across multiple views (Xiong et al., 2025, Jin et al., 2026). Edge changes alter the neighborhoods used for relational propagation, while attribute changes modify both the signals propagated by relational encoders and the neighborhoods of an attribute-derived view. A feature perturbation can influence several representations through different computational paths, producing errors that are correlated across views.

Cross-view alignment (Wang et al., 2021b, Guo et al., 2023) encourages representations of the same node to agree, providing a learning signal across relational and attribute-derived views. Agreement can also coexist with shared errors, and the remaining disagreement may contain both useful view-specific information and corruption effects. The distinction matters because view-unique information can contribute to prediction (Shen et al., 2024), making indiscriminate suppression of disagreement undesirable. We study how regularizing residual interactions complements alignment, focusing on the conditions that allow the regularizer to limit corruption-induced changes.

We examine the geometry of deviations from the mean view representation, using the empirical mean as a common anchor. Centering removes additive components shared by all views, allowing the regularizer to act on relative differences between relational and attribute-derived representations. The cross-Gram matrix of a residual pair measures interactions between feature channels across nodes, retaining information about how the directions of disagreement are related. The resulting feature-channel constraint complements the matched-node comparisons used by contrastive alignment, as depicted in Figure 1.

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.



(a) Agreement alone (b) Residual regularization

Figure 1: Motivation. (a) Agreement can retain shared errors. (b) OCG regularizes relational and semantic residual interactions around the empirical mean. "Cross Gram" denotes interactions between the relational and semantic residuals.

We propose Orthogonal Consensus Graph Learning (OCG), which combines contrastive alignment with a cross-Gram penalty on mean-centered relational and semantic residuals. Schema-induced relational views and an attribute-derived hypergraph provide the inputs for view encoding, followed by refinement, residual regularization, and fusion. Cross-view and within-view contrastive objectives supply the representation-learning signal, and stochastic-path predictive consistency encourages stable predictions under embedding masks.

Our analysis separates the residual penalty into two terms: spectral shrinkage of the semantic residual and its coupling to relational disagreement. We derive a residual-sensitivity bound for corrupted inputs and further connect it to the final fused representation. Experiments on ACM, DBLP, Yelp, and MAG evaluate classification and clustering, with perturbation tests and ablations assessing robustness and the contribution of each component. The main contributions can be summarized as follows:

- We introduce an orthogonality-guided constraint on residual interactions between relational and semantic views, integrated with cross-view alignment and predictive consistency.
- We derive an exact decomposition of the residual objective into coupling and spectral shrinkage, and establish a residual-sensitivity bound that connects to the fused representation.
- We evaluate OCG on four graph benchmarks and assess edge- and feature-perturbation robustness. Component ablations quantify how the regularizer and the other objectives contribute to classification under clean and corrupted inputs.

2 Related Work

Heterogeneous graph learning. Metapath-based methods aggregate information across relational schemas (Wang et al., 2019, Dong et al., 2017, Sun et al., 2011), and multiplex approaches learn from multiple relational views (Mo et al., 2023b, Qian et al., 2024, Shen et al., 2024, Guo et al., 2023). Hypergraph constructions incorporate feature-space neighborhoods to capture attribute-derived relationships (Feng et al., 2019, Jiang et al., 2019, Gao et al., 2023). OCG combines relational and attribute-derived views with an explicit constraint on their residual interactions around a shared mean.

Robust graph learning. Graph structure learning refines topology through learned adjacency or propagation mechanisms (Franceschi et al., 2019, Yu et al., 2021, Jin et al., 2020, Wang et al., 2021a, Wu et al., 2022), while robust representation methods address unreliable inputs through self-supervised objectives (Liu et al., 2022, Li et al., 2022, Zhao et al., 2023, Wong et al., 2026). Closely related multiplex baselines include BTGF (Qian et al., 2024) and InfoMGF (Shen et al., 2024), which models shared and view-unique task information. Our analysis examines residual cross-Gram interactions and the overlap between penalized directions and informative disagreement.

Orthogonality and decorrelation. Ortho-GConv (Guo et al., 2022) regularizes feature transformations to stabilize GNN training, and Barlow Twins (Zbontar et al., 2021) targets an identity cross-correlation matrix between augmented representations. VICReg (Bardes et al., 2022) combines invariance, variance, and within-representation covariance terms to regulate learned representations. The OCG regularizer penalizes full cross-Gram matrices between relational and semantic residuals centered at a multi-view mean.

3 PRELIMINARIES

Notation. We consider M schema-induced relational views $G^{(v)} = (\mathcal{V}, \mathcal{E}^{(v)})$ over a shared node set with $N = |\mathcal{V}|$, adjacency matrices $A^{(v)} \in \{0,1\}^{N\times N}$, and node features $X \in \mathbb{R}^{N\times d_f}$. We also construct an attribute-derived semantic hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E}_h)$ with incidence matrix $H \in \{0,1\}^{N\times|\mathcal{E}_h|}$ (Feng et al., 2019, Yadati et al., 2019). Here "view" refers to a relational schema or a feature-space construction over the same nodes. The notation is summarized in Appendix A.

Consensus and residuals. The empirical consensus anchor is the mean of the encoded view representations, and each view's residual is its deviation from that anchor. Residual orthogonality denotes the cross-Gram penalty on residual interactions defined in Eq. (6).

Problem formulation. We learn representations from a multiplex graph $\mathcal{G} = (\mathcal{V}, \{\mathcal{E}^{(v)}\}_{v=1}^{M}, X)$ using the unsupervised objective in Eq. (9), with labels reserved for the stated downstream evaluation protocol. We study how the residual objective constrains representation changes and assess robustness under edge and node-feature perturbations in Section Section 5.3.

4 METHOD

OCG constructs and encodes schema-induced relational views and an attribute-derived semantic view, then jointly optimizes cross-view alignment, residual regularization, and stochastic-path predictive consistency, as shown in Figure 2.

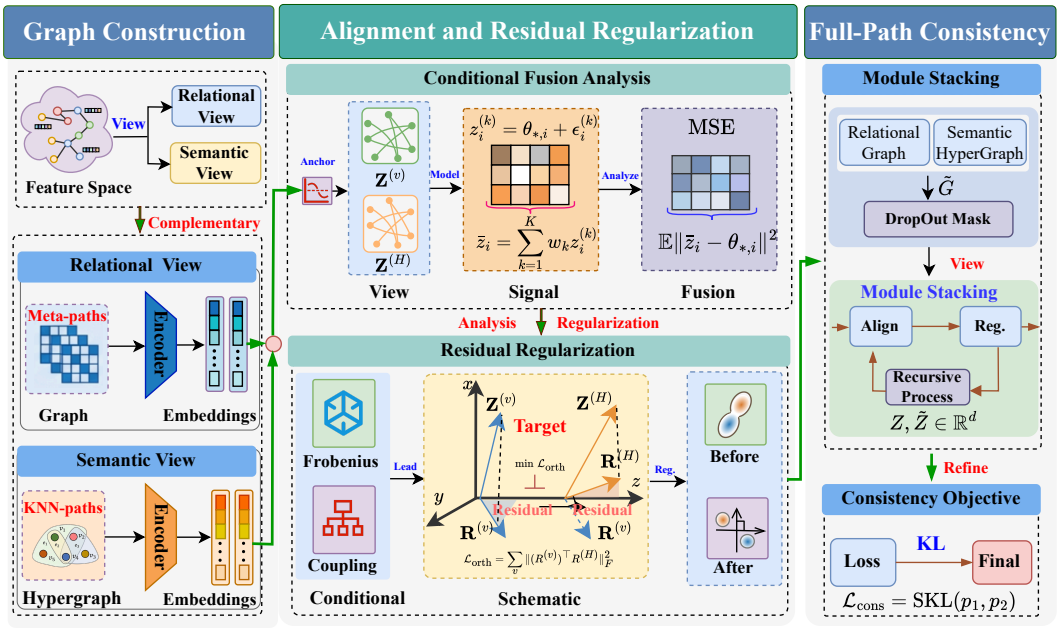


Figure 2: OCG overview. Relational and attribute-derived views are encoded and aligned, with a cross-Gram penalty on their mean-centered residuals and stochastic-path consistency. The upper-middle signal-plus-error model illustrates the conditional averaging analysis. Arrows in the residual schematic illustrate the regularization constraint.

4.1 VIEW ENCODERS

Relational view. We start from M schema-induced relational views with adjacency matrices $\{A^{(v)}\}_{v=1}^{M}$ and a shared feature matrix $X \in \mathbb{R}^{N \times d_f}$. To incorporate multi-hop structural context, we use SGC-style propagation (Wu et al., 2019). Let $\tilde{A}^{(v)} = A^{(v)} + I$ and let $\tilde{D}^{(v)}$ be its degree matrix. With propagation depth r, the view-aware node signals are

$$X^{(v)} = \left[(\tilde{D}^{(v)})^{-1/2} \, \tilde{A}^{(v)} \, (\tilde{D}^{(v)})^{-1/2} \right]^{r} X. \tag{1}$$

Semantic view. We construct an attribute-derived hypergraph to capture feature-space proximity, which may be absent from the observed edges. We call this the semantic view and denote it by $\mathcal{H} = (\mathcal{V}, \mathcal{E}_h)$ with incidence matrix $H \in \{0,1\}^{N \times |\mathcal{E}_h|}$. Each hyperedge connects a node to its k-nearest neighbours in feature space, identified through approximate nearest-neighbour search for scalability (Indyk & Motwani, 1998, Johnson et al., 2021). Clique expansion converts the hypergraph into an additional graph $G^{(H)} = (\mathcal{V}, \mathcal{E}^{(H)})$. The full collection of views is $\{G^{(1)}, \ldots, G^{(M)}, G^{(H)}\}$ with $K = M + 1$.

View-specific encoding. The propagated signals $X^{(v)}$ pass through view-specific feature-weighting learners, whose row-normalized outputs define pairwise similarities. The learned adjacency matrices are obtained through top-k selection, symmetrization, a nonlinearity, and normalization. A shared GCN encoder (Kipf & Welling, 2017) then processes the original features X on each learned view, producing node representations $Z^{(k)} \in \mathbb{R}^{N \times d}$ for $k \in \{1, \ldots, K\}$. We write $z_i^{(k)} \in \mathbb{R}^d$ for the representation of node i under view k. The semantic embedding passes through a linear projection and batch normalization, then row normalization, before refinement.

4.2 REFINEMENT AND FUSION

The encoded views are refined before computing the residual penalty and the final readout. Let $Z_{(0)}^{(k)}$ denote the encoded view matrices and $\bar{Z}_{(t)}$ their mean across all K views at block t. For $t = 0, \ldots, T-1$, each view is first combined with this mean to form a candidate representation:

$$U_{(t)}^{(k)} = \mathrm{MLP}_k\big((\bar{Z}_{(t)} + Z_{(t)}^{(k)})/2\big). \tag{2}$$

The update blends this candidate with the current representation:

$$Z_{(t+1)}^{(k)} = (1-\alpha) Z_{(t)}^{(k)} + \alpha U_{(t)}^{(k)}. \tag{3}$$

Each view has a two-layer MLP whose parameters are shared across its refinement blocks. The coefficient $\alpha \in [0,1]$ controls the update, and $T = 0$ disables refinement. The residual loss is evaluated on the final refined matrices $Z_{(T)}^{(k)}$ before fusion normalization, and unindexed view representations below refer to those matrices. After refinement, a learned scalar gate $g_H = \sigma(a_H)$ weights the semantic view in the final readout:

$$\tilde{Z} = \frac{\sum_{v=1}^{M} \nu(Z_{(T)}^{(v)}) + g_H \nu(Z_{(T)}^{(H)})}{M + g_H}, \tag{4}$$

where ν normalizes each row as $z / \max(\|z\|_2, \varepsilon)$, with a numerical floor $\varepsilon > 0$. The gate is constant when a trained model is evaluated on different inputs. Setting $g_H = 1$ gives the mean of the normalized views, which need not equal the mean used in the residual loss.

4.3 RESIDUAL REGULARIZATION

The residual penalty complements alignment by constraining interactions among view-specific deviations from a shared anchor. We define a shared anchor as the mean of the final refined representations across all K views:

$$\bar{Z} = \frac{1}{K} \left(\sum_{v=1}^{M} Z^{(v)} + Z^{(H)} \right). \tag{5}$$

The residuals are $R^{(v)} = Z^{(v)} - \bar{Z}$ and $R^{(H)} = Z^{(H)} - \bar{Z}$. Our orthogonality regularizer penalizes cross-Gram interactions between relational and semantic residuals:

$$\mathcal{L}_{\text{orth}} = \sum_{v=1}^{M} \left\| (R^{(v)})^\top R^{(H)} \right\|_F^2. \tag{6}$$

The penalty measures interactions between residual feature channels across nodes. Mean centering makes the residuals sum to zero, coupling their behavior under alignment and regularization. Proposition 1 gives an exact decomposition into spectral shrinkage of the semantic residual and its coupling to relational disagreement. Each penalized pair contains one relational view and the semantic view, constraining interactions between schema-derived and attribute-derived representations.

4.4 Training objective

Cross-view alignment. We align relational and semantic views with a symmetric InfoNCE objective (van den Oord et al., 2018, Wang et al., 2021b). Let $s_{ij}^{vH} = \text{sim}(z_i^{(v)}, z_j^{(H)})/\tau$, where sim is cosine similarity and $\tau > 0$ is the temperature. The directed loss is

$$\ell_{v \to H} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s_{ii}^{vH})}{\sum_{j=1}^{N} \exp(s_{ij}^{vH})}. \tag{7}$$

We average the forward and reverse losses to obtain a symmetric loss for each view pair, then average over relational views to form $\mathcal{L}_{\text{ssl}}^{\text{cross}}$. Applying the same symmetric loss between each view and its stochastic augmentation gives $\mathcal{L}_{\text{ssl}}^{\text{aug}}$. The self-supervised objective $\mathcal{L}_{\text{ssl}}$ is the sum of the cross-view and augmentation terms. Additional pairings used in the controlled experiments are specified in Appendix B.

Predictive consistency. To encourage prediction consistency across stochastic paths, we independently mask and fuse the view embeddings before applying an auxiliary prediction head. Let $p_\theta(\cdot \mid \mathcal{G}, i; \xi)$ denote its softmax output under mask ξ. The auxiliary head is trained through the consistency loss, with a separate classifier used for downstream evaluation. Writing $p_i^{(s)} = p_\theta(\cdot \mid \mathcal{G}, i; \xi_s)$ for two independent masks ξ_1, ξ_2, we penalize their disagreement with a bidirectional KL divergence (Liang et al., 2021):

$$\mathcal{L}_{\text{cons}} = \frac{1}{2N} \sum_{i=1}^{N} \left[\text{KL}\big(p_i^{(1)} \,\|\, p_i^{(2)}\big) + \text{KL}\big(p_i^{(2)} \,\|\, p_i^{(1)}\big) \right]. \tag{8}$$

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{ssl}} + \lambda_o \mathcal{L}_{\text{orth}} + \lambda_c \mathcal{L}_{\text{cons}}, \tag{9}$$

where $\lambda_o, \lambda_c \geq 0$ weight residual regularization and predictive consistency, respectively.

4.5 Residual analysis

We analyze the representations at the block where Eq. (6) is evaluated, before the final row normalization used for fusion. Let $M, N, d \geq 1$, write $\|U\|_{\text{rms}} = \|U\|_F / \sqrt{N}$, and use the scaled loss $\ell_o = \mathcal{L}_{\text{orth}} / N^2$ for analysis, with $\mathcal{L}_{\text{orth}}$ retained as the training loss.

Residual geometry. Let $Z_R = M^{-1} \sum_v Z^{(v)}$, $Q_v = Z^{(v)} - Z_R$, and $h = R^{(H)}$. Since $\sum_v Q_v = 0$ and $R^{(v)} = Q_v - h/M$, expanding the loss gives the following characterization.

Proposition 1 (Residual geometry)**.** *Adding the same matrix to every view leaves $\mathcal{L}_{\text{orth}}$ unchanged. The loss decomposes as*

$$\mathcal{L}_{\text{orth}} = \sum_{v=1}^{M} \|Q_v^\top h\|_F^2 + \frac{1}{M} \|h^\top h\|_F^2, \tag{10}$$

and $\|h\|_{\text{rms}} \leq (M d \ell_o)^{1/4}$. The loss is zero exactly when $Z^{(H)} = Z_R$.

The coupling term measures how h interacts with relational disagreement, while the spectral term penalizes the fourth powers of its singular values. Common signal and common error cancel through mean centering, leaving the penalty dependent on relative view differences. The proof and an example distinguishing this penalty from squared-distance alignment are given in Appendix C.1.

Perturbation control. Evaluate clean and corrupted inputs with the same trained model, and write $\Delta h = h_c - h_0$. Using the corrupted relational deviations, define

$$B_c = \frac{1}{N} \sum_v Q_{v,c} Q_{v,c}^\top \succeq 0, \tag{11}$$

with P_κ the spectral projector onto eigenvalues at least $\kappa > 0$, acting in node-sample space for the analysis.

Assumption 1 (Directional separation)**.** *For some $b_s, b_n \geq 0$, the clean residual and its perturbation satisfy*

$$\|P_\kappa h_0\|_{\mathrm{rms}} \leq b_s, \qquad \|(I - P_\kappa)\Delta h\|_{\mathrm{rms}} \leq b_n. \tag{12}$$

The condition is useful when strongly penalized directions contain little clean residual information and capture most of the corruption-induced change.

Theorem 1 (Conditional residual sensitivity)**.** *Under Assumption 1,*

$$\|\Delta h\|_{\mathrm{rms}} \leq D_c := \left[(\sqrt{\ell_{o,c}/\kappa} + b_s)^2 + b_n^2\right]^{1/2}, \tag{13}$$

where $\ell_{o,c}$ is evaluated on the corrupted representations.

Proof sketch. Equation (10) bounds the RMS norm of $P_\kappa h_c$ by $\sqrt{\ell_{o,c}/\kappa}$. Apply the triangle inequality to the projected change, then combine the selected and complementary components of Δh using their orthogonality. □

From residuals to fusion. Let $g \in [0, 1]$ be the learned scalar gate, held fixed when comparing clean and corrupted inputs. Denote by F_g the weighted average of the unnormalized views, using the weights in Eq. (4), and by E_g the correction from normalizing each view. Then

$$\tilde{Z} = \bar{Z} - \gamma(g)h + E_g, \qquad \gamma(g) = \frac{1-g}{M+g}.$$

The fusion identity separates contributions from the common anchor, the semantic residual, and view normalization, allowing their changes to be bounded individually.

Corollary 1 (Fusion sensitivity)**.** *Let a_c and e_c denote the RMS changes in the common anchor and the normalization correction, respectively. Under the theorem's conditions,*

$$\|\tilde{Z}_c - \tilde{Z}_0\|_{\mathrm{rms}} \leq a_c + e_c + \gamma(g) D_c. \tag{14}$$

The D_c term combines the corrupted-input residual loss with the directional separation bounds, while a_c and e_c account for anchor drift and normalization effects. The remaining proofs, an extension to varying gates, and a conditional signal-plus-error averaging analysis are given in Appendices C.2 to C.5.

5 Experiments

We evaluate the learned representations through node classification and clustering, test robustness to input corruption, and measure the contribution of each objective through ablation. A controlled experiment examines the two terms in the residual-loss decomposition.

5.1 Setup

Datasets and metrics. We evaluate OCG on four heterogeneous multiplex benchmarks: ACM, DBLP (Yun et al., 2019), Yelp (Lu et al., 2019), and MAG (Wang et al., 2020). Table 1 summarizes their sizes, splits, and relational views. Dataset provenance is described in Appendix D.2. Performance is measured by Macro-F_1 and Micro-F_1 for node classification, and by ACC, NMI, F1, and ARI for K-means node clustering.

Table 1: Dataset statistics.

Dataset	Nodes	Features	Classes	Train/Val/Test	Edges (Relation)
ACM	3,025	1,902	3	600 / 300 / 2,125	26,416 (PAP) 2,197,556 (PSP)
DBLP	2,957	334	4	600 / 300 / 2,057	2,398 (APA) 1,460,724 (APCPA)
Yelp	2,614	82	3	300 / 300 / 2,014	525,718 (BUB) 2,475,108 (BSB) 1,484,692 (BLB)
MAG	113,919	128	4	40,000 / 10,000 / 63,919	1,806,596 (PP) 10,067,799 (PAP)

Table 2: Node classification results ($\% \pm \sigma$). OOM denotes out of memory under the evaluated configuration.

Method	ACM		DBLP		Yelp		MAG	
	Macro-F_1	Micro-F_1	Macro-F_1	Micro-F_1	Macro-F_1	Micro-F_1	Macro-F_1	Micro-F_1
GCN	90.27±0.59	90.18±0.61	90.01±0.32	90.99±0.28	78.01±1.89	81.03±1.81	75.98±0.07	75.76±0.10
GAT	91.52±0.62	91.46±0.62	90.22±0.37	91.13±0.40	82.12±1.47	84.43±1.56	OOM	
HAN	91.67±0.39	91.47±0.22	90.53±0.24	91.47±0.22	88.49±1.73	88.78±1.40	OOM	
HGNN+	88.53±0.24	88.38±0.37	90.39±0.24	91.44±0.43	91.76±0.25	91.49±0.46	OOM	
LDS	92.35±0.43	92.05±0.26	88.11±0.86	88.74±0.85	75.98±2.35	78.14±1.98	OOM	
GRCN	93.04±0.17	92.94±0.18	88.33±0.47	89.43±0.44	76.05±1.05	80.68±0.96	OOM	
IDGL	91.69±1.24	91.63±1.24	89.65±0.60	90.61±0.56	76.98±5.78	79.15±5.06	OOM	
ProGNN	90.57±1.03	90.50±1.29	83.13±1.56	84.83±1.36	51.76±1.46	58.39±1.25	OOM	
GEN	87.91±2.78	87.88±2.61	89.74±0.69	90.65±0.71	80.43±3.78	82.68±2.84	OOM	
NodeFormer	91.33±0.77	90.60±0.95	79.54±0.78	80.56±0.62	91.69±0.65	90.59±1.21	77.21±0.18	77.08±0.19
SUBLIME	92.42±0.16	92.13±0.37	90.98±0.37	91.82±0.27	79.68±0.79	82.99±0.82	75.96±0.05	75.71±0.03
STABLE	83.54±4.20	83.38±4.51	75.18±1.95	76.42±1.95	71.48±4.71	76.62±2.75	OOM	
GSR	92.14±1.08	92.11±0.99	76.59±0.45	77.69±0.42	83.85±0.76	85.73±0.54	OOM	
MGFNN	88.66±0.29	88.47±0.34	89.55±0.31	90.67±0.45	74.59±0.53	77.86±0.62	66.40±0.47	66.01±0.34
HDMI	91.01±0.32	90.86±0.31	89.91±0.49	90.89±0.51	80.73±0.64	84.05±0.91	72.22±0.14	71.84±0.15
DMG	90.42±0.36	90.31±0.35	90.42±0.57	91.34±0.49	91.61±0.62	90.24±0.81	76.34±0.09	76.13±0.10
BTGF	91.75±0.11	91.62±0.11	90.71±0.24	91.57±0.21	92.81±1.12	91.37±1.28	OOM	
InfoMGF	93.21±0.22	93.14±0.21	90.99±0.36	91.93±0.29	93.09±0.27	92.02±0.34	77.25±0.06	77.11±0.06
OCG	**93.64±0.35**	**93.53±0.28**	**91.63±0.29**	**92.53±0.31**	**93.55±0.19**	**92.57±0.38**	**77.58±0.07**	**77.52±0.18**

Baselines. We compare against four families: (1) *Basic GNNs*: GCN (Kipf & Welling, 2017), GAT (Veličković et al., 2018), HAN (Wang et al., 2019), HGNN+ (Gao et al., 2023). (2) *Graph Structure Learning*: LDS (Franceschi et al., 2019), GRCN (Yu et al., 2021), ProGNN (Jin et al., 2020), IDGL (Chen et al., 2020), GEN (Wang et al., 2021a), NodeFormer (Wu et al., 2022). (3) *Robust Learning*: SUBLIME (Liu et al., 2022), STABLE (Li et al., 2022), GSR (Zhao et al., 2023). (4) *Multiplex Learning*: HDMI (Jing et al., 2021), DMG (Mo et al., 2023b), BTGF (Qian et al., 2024), InfoMGF (Shen et al., 2024), MGFNN (Liu et al., 2026).

Evaluation protocol. We first train the encoder with Eq. (9) without labels, then freeze the embeddings and train a linear classifier on the training split. Validation data are used for hyperparameter selection, and classification scores are measured on the test split. The benchmark encoder uses Adam with learning rate 10^{-3} and weight decay 5×10^{-4}, for up to 1000 epochs with patience 50 on validation $\mathcal{L}_{\text{ssl}}$. The default embedding dimension is 512, with two refinement blocks. The component experiment uses a separate controlled configuration. Robustness tests follow the edge and feature perturbation protocols in Appendix D.4.

5.2 Main results

As shown in Table 2, OCG achieves the highest mean Macro-F1 and Micro-F1 among the evaluated methods on ACM, DBLP, Yelp, and MAG. Compared with InfoMGF, OCG improves Macro-F1 by

Table 3: Robustness analysis on ACM and DBLP. *Edge Ptb*: random edge addition/deletion at rate η_{edge} on all relational views simultaneously. *Node Ptb*: ℓ_∞-bounded transfer PGD feature attack at rate η_{node}. We report Macro-F1 and Micro-F1 ($\% \pm \sigma$).

Dataset	Method	*Edge Ptb*				*Node Ptb*			
		$\eta_{edge}=0.1$		$\eta_{edge}=0.2$		$\eta_{node}=0.1$		$\eta_{node}=0.2$	
		Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
ACM	GCN	81.79±0.20	82.84±0.19	76.33±0.28	75.79±0.24	65.29±0.18	67.89±0.20	50.89±0.38	59.80±0.41
	HAN	78.72±0.25	78.31±0.18	73.42±0.34	73.51±0.31	78.83±0.14	79.30±0.12	63.98±0.20	65.11±0.26
	HGNN+	82.08±0.31	82.21±0.35	83.41±0.48	85.74±0.36	83.69±0.16	84.04±0.15	66.58±0.23	70.36±0.12
	SUBLIME	79.71±0.34	79.19±0.40	78.28±0.29	77.92±0.31	80.45±0.38	80.75±0.22	58.23±0.31	60.18±0.24
	NodeFormer	74.62±0.60	73.48±0.49	72.48±0.45	71.20±0.62	86.01±0.34	86.07±0.31	67.68±0.56	70.11±0.31
	MGFNN	81.78±0.42	82.02±0.45	74.69±0.43	74.18±0.34	72.65±0.46	72.52±0.43	70.39±0.31	70.23±0.20
	InfoMGF	87.32±0.21	87.32±0.25	86.14±0.30	85.89±0.27	82.51±0.39	82.68±0.13	65.51±0.25	66.26±0.21
	OCG	**89.77±0.26**	**89.68±0.31**	**87.49±0.26**	**88.33±0.33**	**87.66±0.24**	**87.70±0.20**	**71.23±0.17**	**70.68±0.19**
DBLP	GCN	85.03±0.44	86.78±0.39	81.05±0.36	82.64±0.32	70.81±0.18	71.75±0.45	63.91±0.33	65.02±0.19
	HAN	86.75±0.26	86.92±0.35	84.91±0.26	85.73±0.40	75.23±0.13	76.39±0.14	65.40±0.16	64.56±0.18
	HGNN+	86.26±0.21	86.01±0.34	83.67±0.24	85.34±0.21	74.98±0.14	74.93±0.13	64.09±0.12	64.37±0.30
	SUBLIME	84.67±0.36	85.95±0.26	79.47±0.33	82.55±0.28	71.23±0.31	70.54±0.32	58.87±0.32	56.68±0.24
	NodeFormer	73.48±0.63	74.62±0.54	71.20±0.38	72.48±0.61	75.46±0.26	75.06±0.35	65.85±0.47	64.83±0.42
	MGFNN	82.49±0.34	84.54±0.35	79.89±0.41	82.35±0.44	75.79±0.26	75.39±0.33	58.78±0.38	61.11±0.43
	InfoMGF	86.86±0.40	86.12±0.32	84.80±0.35	85.65±0.43	73.43±0.25	75.91±0.15	59.07±0.29	59.45±0.15
	OCG	**88.51±0.34**	**87.84±0.28**	**86.52±0.29**	**86.32±0.33**	**76.17±0.29**	**76.45±0.17**	**66.08±0.19**	**65.19±0.28**

0.43, 0.64, 0.46, and 0.33 percentage points, respectively, with corresponding Micro-F1 gains of 0.39, 0.60, 0.55, and 0.41 points. The improvements occur across datasets with different graph sizes, feature dimensions, and relational structures, with DBLP showing the largest gain on both metrics. We further assess the learned representations through clustering, where OCG leads all four metrics on ACM, DBLP, and MAG and achieves the highest ARI on Yelp. Detailed clustering results are provided in Appendix E.1.

5.3 Robustness

The robustness results in Table 3 show that OCG retains its classification advantage under input corruption, achieving the highest mean Macro-F1 and Micro-F1 on ACM and DBLP at every tested perturbation rate. On ACM, increasing η_{edge} from 0.1 to 0.2 reduces OCG Macro-F1 from 89.77 to 87.49, while its margins over InfoMGF remain 2.45 and 1.35 percentage points at the respective rates. Under the feature attack at $\eta_{node} = 0.2$, OCG reaches 71.23 Macro-F1, exceeding InfoMGF by 5.72 points and the strongest baseline in that column, MGFNN, by 0.84 points. The results on DBLP follow the same pattern, with OCG reaching 86.52 Macro-F1 under edge perturbation at $\eta_{edge} = 0.2$ and 66.08 under the feature attack at $\eta_{node} = 0.2$. The consistent ranking across both datasets supports OCG's effectiveness under the evaluated structural and feature perturbations.

5.4 Ablations

We ablate OCG on ACM under clean, *Edge Ptb* ($\eta_{edge} = 0.2$), and *Node Ptb* ($\eta_{node} = 0.2$) settings, with variants isolating objective and architectural components.

Objective variants. (i) *w/o* $\mathcal{L}_{\text{ssl}}^{\text{cross}}$: removes cross-view contrastive alignment between relational and semantic views. (ii) *w/o* $\mathcal{L}_{\text{ssl}}^{\text{aug}}$: removes the augmented-view contrastive term. (iii) *w/o* $\mathcal{L}_{\text{orth}}$: removes the residual cross-Gram regularizer. (iv) *w/o* $\mathcal{L}_{\text{cons}}$: removes predictive consistency.

Architectural variants. (v) *w/o* $\mathcal{H}$: removes the semantic hypergraph view ($K = M$). (vi) *w/o Aug*: disables feature-level augmentation. (vii) *w/o Stacking*: sets $T = 0$ in Eq. (3), removing recursive refinement.

Table 4 shows that removing $\mathcal{L}_{\text{orth}}$ reduces Macro-F1 by 5.33, 0.93, and 24.74 percentage points on clean, edge-perturbed, and feature-perturbed ACM, respectively. Removing the semantic view reduces feature-attack Macro-F1 by 11.20 points and edge-perturbation Macro-F1 by 0.39 points. Removing predictive consistency reduces feature-attack Macro-F1 by 4.53 points.

Table 4: Ablation on ACM. Upper: objective components. Lower: architectural components. Macro-F1 and Micro-F1 ($\% \pm \sigma$). *Edge Ptb*: structural perturbation at rate η_{edge}. *Node Ptb*: transfer PGD feature attack at rate η_{node}.

	Variant	*Clean*		*Edge Ptb* ($\eta_{edge}=0.2$)		*Node Ptb* ($\eta_{node}=0.2$)	
		Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Objective	w/o $\mathcal{L}_{\text{ssl}}^{\text{cross}}$	92.24±0.19	92.31±0.23	87.24±0.32	87.58±0.34	68.68±0.46	69.79±0.31
	w/o $\mathcal{L}_{\text{ssl}}^{\text{aug}}$	91.86±0.24	91.56±0.21	87.35±0.52	87.73±0.55	64.32±0.53	66.15±0.40
	w/o $\mathcal{L}_{\text{orth}}$	88.31±0.38	88.92±0.47	86.56±0.23	87.21±0.21	46.49±0.35	53.15±0.22
	w/o $\mathcal{L}_{\text{cons}}$	92.54±0.27	92.42±0.23	87.24±0.38	87.68±0.27	66.70±0.43	67.41±0.36
Arch.	w/o $\mathcal{H}$	91.99±0.34	92.21±0.38	87.10±0.29	86.36±0.31	60.03±0.27	65.02±0.25
	w/o Aug	93.04±0.17	92.89±0.15	86.47±0.51	86.93±0.87	66.73±0.28	68.31±0.34
	w/o Stacking	92.71±0.28	92.57±0.27	87.13±0.21	87.46±0.29	66.64±0.20	68.05±0.27
	OCG	**93.64±0.35**	**93.53±0.28**	**87.49±0.26**	**88.33±0.33**	**71.23±0.17**	**70.68±0.19**

Removing cross-view alignment lowers Macro-F1 by 0.25 points under edge perturbation and 2.55 points under the feature attack. The variants without feature-level augmentation and without stacking each score lower on both metrics across all three settings. Under edge perturbation, the respective Macro-F1 drops are 1.02 and 0.36 points. Differences from related regularizers are discussed in Appendix E.2.

Residual-loss components. A separate ACM experiment compares the coupling and spectral terms in Eq. (10). Spectral shrinkage accounts for much of the reduction in residual interaction and magnitude, but lower residual interaction does not consistently improve classification. The experiment uses a smaller model without refinement and three random seeds. The setup and results are detailed in Appendix F.

6 Conclusion

We introduced OCG, a heterogeneous graph representation framework that combines cross-view alignment with residual interaction regularization and stochastic-path predictive consistency. By decomposing the residual objective into coupling and spectral shrinkage, OCG provides a geometric characterization of the interaction between relational disagreement and semantic residuals, and links residual sensitivity to the final fused representation. Across ACM, DBLP, Yelp, and MAG, OCG achieves strong classification performance and remains robust under structural and feature perturbations. Ablation and controlled studies further show that residual regularization is a central contributor to robustness, with the coupling and spectral components playing complementary roles.

Reproducibility Statement

The architecture, learning objectives, and training procedure are described in Section 4. Additional details required for reproduction are provided in the appendices, including notation, implementation and hyperparameter settings, dataset and preprocessing information, evaluation protocols, and robustness configurations. We also include complete proofs of the theoretical results together with additional experimental analyses. For the controlled experiments, the random seeds, model configurations, and aggregation conventions are explicitly reported. Together, these materials document the settings used for the main experiments and the auxiliary studies. See Appendices A to F for details.

AI use statement

Generative AI tools were used to assist with language polishing and the verification of mathematical arguments. The authors are responsible for the final manuscript, mathematical claims, and reported results.

REFERENCES

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pp. 1247–1255. PMLR, 2013.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I del Genio, Jesús Gómez-Gardeñes, Miguel Romance, Irene Sendiña-Nadal, Zhen Wang, and Massimiliano Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

Yu Chen, Lingfei Wu, and Mohammed Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 135–144, 2017.

Shaohua Fan, Xiao Wang, Chuan Shi, Emiao Lu, Ken Lin, and Bai Wang. One2multi graph autoencoder for multi-view graph clustering. In *proceedings of the web conference 2020*, pp. 3070–3076, 2020.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1972–1982. PMLR, 2019.

Yue Gao, Yifan Feng, Shuyi Ji, and Rongrong Ji. HGNN$^+$: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3181–3199, 2023. doi: 10.1109/TPAMI.2022.3182052.

Kai Guo, Kaixiong Zhou, Xia Hu, Yu Li, Yi Chang, and Xin Wang. Orthogonal graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3996–4004, 2022. doi: 10.1609/aaai.v36i4.20316.

Yuanyuan Guo, Yu Xia, Rui Wang, Rongcheng Duan, Lu Li, and Jiangmeng Li. M2HGCL: Multi-scale meta-path integrated heterogeneous graph contrastive learning. In *International Conference on Advanced Data Mining and Applications*, pp. 352–367. Springer, 2023.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*. ACM, 1998.

Jianwen Jiang, Yuxuan Wei, Yifan Feng, Jingxuan Cao, and Yue Gao. Dynamic hypergraph neural networks. In *Ijcai*, pp. 2635–2641, 2019.

Di Jin, Haotian Zhao, Xiaobao Wang, Fengyu Yan, and Dongxiao He. Mitigating noise and imbalance in social governance graphs for multi-type risk assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14973–14981, 2026.

Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. Graph structure learning for robust graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 66–74, 2020.

Baoyu Jing, Chanyoung Park, and Hanghang Tong. Hdmi: High-order deep multiplex infomax. In *Proceedings of the web conference 2021*, pp. 2414–2424, 2021.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

Thomas N. Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016. URL https://arxiv.org/abs/1611.07308.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

Kuan Li, Yang Liu, Xiang Ao, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. Reliable representations make a stronger defender: Unsupervised structure refinement for robust gnn. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 925–935, 2022.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10890–10905, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5a66b9200f29ac3fa0ae244cc2a51b39-Abstract.html.

Zhiping Lin and Zhao Kang. Graph filter-based multi-view attributed graph clustering. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2723–2729, 2021. doi: 10.24963/ijcai.2021/375. URL https://www.ijcai.org/proceedings/2021/375.

Yixin Liu, Yu Zheng, Daokun Zhang, Hongxu Chen, Hao Peng, and Shirui Pan. Towards unsupervised deep graph structure learning. In *Proceedings of the ACM web conference 2022*, pp. 1392–1403, 2022.

Yunhui Liu, Zhen Tao, Xiang Zhao, Jianhua Zhao, Tao Zheng, and Tieke He. Learning accurate, efficient, and interpretable mlps on multiplex graphs via node-wise multi-view ensemble distillation. In *Database Systems for Advanced Applications: DASFAA 2025*, pp. 120–135. Springer, 2026. doi: 10.1007/978-981-95-3906-2_8.

Yuanfu Lu, Chuan Shi, Linmei Hu, and Zhiyuan Liu. Relation structure-aware heterogeneous information network embedding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4456–4463, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Yujie Mo, Yuhuan Chen, Yajie Lei, Liang Peng, Xiaoshuang Shi, Changan Yuan, and Xiaofeng Zhu. Multiplex graph representation learning via dual correlation reduction. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12814–12827, 2023a.

Yujie Mo, Yajie Lei, Jialie Shen, Xiaoshuang Shi, Heng Tao Shen, and Xiaofeng Zhu. Disentangled multiplex graph representation learning. In *International conference on machine learning*, pp. 24983–25005. PMLR, 2023b.

ErLin Pan and Zhao Kang. Multi-view contrastive graph clustering. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Xiaowei Qian, Bingheng Li, and Zhao Kang. Upper bounding barlow twins: A novel filter for multi-relational clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 14660–14668, 2024.

Zhixiang Shen, Shuo Wang, and Zhao Kang. Beyond redundancy: Information-aware unsupervised multiplex graph structure learning. *Advances in neural information processing systems*, 37: 31629–31658, 2024.

Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. PathSim: Meta path-based top-K similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4 (11):992–1003, 2011.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ. accepted as poster.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1): 396–413, 2020.

Ruijia Wang, Shuai Mou, Xiao Wang, Wanpeng Xiao, Qi Ju, Chuan Shi, and Xing Xie. Graph structure estimation neural networks. In *Proceedings of the web conference 2021*, pp. 342–353, 2021a.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. Heterogeneous graph attention network. In *The World Wide Web Conference (WWW)*, 2019.

Xiao Wang, Nian Liu, Hui Han, and Chuan Shi. Self-supervised heterogeneous graph neural network with co-contrastive learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1726–1736, 2021b.

Zhen Hao Wong, Hansi Yang, Quanming Yao, and Yaqing Wang. Robust heterogeneous network representation learning by multifaceted curriculum training. *Neural Networks*, 196:108438, 2026.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Q Weinberger. Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019.

Qitian Wu, Wentao Zhao, Zenan Li, David P Wipf, and Junchi Yan. Nodeformer: A scalable graph structure learning transformer for node classification. *Advances in neural information processing systems*, 35:27387–27401, 2022.

Zhang Xiong, Cheng Xie, Haoran Duan, and Beibei Yu. NoiseHGNN: Synthesized similarity graph-based neural network for noised heterogeneous graph representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21725–21733, 2025.

Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. Hypergcn: A new method for training graph convolutional networks on hypergraphs. In *Advances in Neural Information Processing Systems*, 2019.

Donghan Yu, Ruohong Zhang, Zhengbao Jiang, Yuexin Wu, and Yiming Yang. Graph-revised convolutional network. In *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2020*, pp. 378–393. Springer, 2021. doi: 10.1007/978-3-030-67664-3_23.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. *Advances in neural information processing systems*, 32, 2019.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 793–803, 2019.

Jianan Zhao, Qianlong Wen, Mingxuan Ju, Chuxu Zhang, and Yanfang Ye. Self-supervised graph structure refinement for graph neural networks. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 159–167, 2023. doi: 10.1145/3539597.3570455. URL https://github.com/AndyJZhao/WSDM23-GSR.

A NOTATION

Table 5 summarizes the notation used throughout the paper.

Table 5: Main notation.

Notation	Description
$\mathcal{V}$, N	Node set and its cardinality.
$A^{(v)} \in \{0,1\}^{N\times N}$	Adjacency matrix of the v-th relational view.
$X \in \mathbb{R}^{N\times d_f}$	Shared node feature matrix.
M, $K = M+1$	Number of relational views / total views.
$\mathcal{H} = (\mathcal{V}, \mathcal{E}_h)$	Semantic hypergraph view.
$H \in \{0,1\}^{N\times\lvert\mathcal{E}_h\rvert}$	Hypergraph incidence matrix.
r, k, d	SGC depth / KNN neighbours / hidden dimension.
$Z^{(k)} \in \mathbb{R}^{N\times d}$	Node representations for view k.
$\bar{Z}$	Shared anchor (mean across K views).
$R^{(v)} = Z^{(v)} - \bar{Z}$	View-specific residual.
$\theta_{*,i}$, $\epsilon_i^{(k)}$	Ideal signal / view-specific error of node i.
σ_k^2	Error variance of view k.
$\hat{C}_{v,\ell} = \frac{1}{N}(R^{(v)})^\top R^{(\ell)}$	Scaled empirical residual cross-Gram matrix.
$\mathcal{L}_{\text{ssl}} = \mathcal{L}_{\text{ssl}}^{\text{cross}} + \mathcal{L}_{\text{ssl}}^{\text{aug}}$	Cross-view and augmented-view contrastive loss.
$\mathcal{L}_{\text{orth}}$, λ_o	Residual cross-Gram loss and its weight.
$\mathcal{L}_{\text{cons}}$, λ_c	Predictive consistency loss and its weight.
$\mathcal{L}_{\text{pair}}$	Mean pairwise squared distance in the latent-error model.
η_{edge}, η_{node}	Edge perturbation / adversarial attack rate.
$g_H \in (0,1)$	Learned semantic-view fusion weight.

B IMPLEMENTATION DETAILS

Section 4 defines view encoding, refinement, fusion, and the training losses. Here we specify the additional contrastive pairings in the controlled implementation. A further structure learner takes concatenated view signals to construct a fused graph. Both original and augmented view embeddings undergo refinement before contrastive losses are computed.

Contrastive pairings. The controlled experiments add contrastive terms for every unordered pair of views, along with terms aligning each view to an embedding Z_F from the learned fused graph. Writing $\ell(U, V)$ for symmetric InfoNCE after the appropriate projection, their cross-view term is

$$\mathcal{L}_{\text{ssl}}^{\text{cross}} = \frac{\sum_{k<l} w_{kl}\ell(Z^{(k)}, Z^{(l)})}{\sum_{k<l} w_{kl}} + \frac{\sum_k w_k \ell(Z_F, Z^{(k)})}{\sum_k w_k}. \tag{15}$$

The augmentation term averages $w_k\ell(Z^{(k)}, Z_{\text{aug}}^{(k)})$ with the same denominator. Semantic-view terms receive the learned semantic weight, and other terms have unit weight. Z_F denotes the contrastive embedding of the learned fused graph, and $\tilde{Z}$ in Eq. (4) denotes the view-average readout. All contrastive pairings remain fixed across the exact-component comparisons. Masking, fusion, and temperature settings for the linear consistency head with dropout are given in Appendix F.1.

C PROOFS AND SCOPE

We first prove the deterministic results for the representations used in the residual loss, then account for row normalization in the fusion head. The final two subsections analyze averaging under a signal-plus-error model and relate that model to the observable residuals. Throughout, $M, N, d \geq 1$, $K = M + 1$, and $\ell_o = \mathcal{L}_{\text{orth}}/N^2$.

C.1 RESIDUAL GEOMETRY

Proof of Proposition 1. Adding a common matrix B to every view also adds B to the mean anchor, so every residual is unchanged.

To separate disagreement among relational views from their disagreement with the semantic view, put $D = Z^{(H)} - Z_R$. Subtracting the mean anchor gives

$$\bar{Z} = Z_R + \frac{D}{M+1}, \qquad h = \frac{M}{M+1}D, \qquad R^{(v)} = Q_v - \frac{h}{M}, \qquad \sum_v Q_v = 0.$$

Writing $G_h = h^\top h$ and expanding the squared Frobenius norms yields

$$\begin{aligned}\mathcal{L}_{\text{orth}} &= \sum_v \|Q_v^\top h - G_h/M\|_F^2 \\ &= \sum_v \|Q_v^\top h\|_F^2 - \frac{2}{M}\left\langle \sum_v Q_v^\top h, G_h \right\rangle_F + \frac{1}{M}\|G_h\|_F^2 \\ &= \sum_v \|Q_v^\top h\|_F^2 + \frac{1}{M}\|h^\top h\|_F^2.\end{aligned}$$

The mixed term vanishes because the Q_v sum to zero. The remaining quartic term controls the size of h: if s_j are its singular values, there are at most d nonzero values, so Cauchy-Schwarz gives

$$\|h^\top h\|_F^2 = \sum_j s_j^4 \geq \frac{1}{d}\left(\sum_j s_j^2\right)^2 = \frac{1}{d}\|h\|_F^4.$$

Hence $\mathcal{L}_{\text{orth}} \geq \|h\|_F^4/(Md)$, which gives the RMS bound after dividing by N^2 and taking fourth roots. Zero loss requires $h = 0$, and the converse follows directly from the definition of the loss. Combining the equivalence with $h = M(Z^{(H)} - Z_R)/(M+1)$ gives the stated zero-loss condition. □

An upper bound complements this lower bound:

$$\frac{\|h\|_F^4}{Md} \leq \mathcal{L}_{\text{orth}} \leq \left(\sum_v \|R^{(v)}\|_F^2\right)\|h\|_F^2. \tag{16}$$

The upper bound follows by applying Cauchy-Schwarz to each entry of $U^\top V$ and summing. Zero loss permits distinct relational views whose mean equals the semantic view.

Lemma 1 (Residual interaction bound)**.** *The observable residuals satisfy*

$$\sum_{v=1}^M \left|\frac{1}{N}\sum_i \langle r_i^{(v)}, r_i^{(H)}\rangle\right| \leq \sqrt{dM\ell_o}. \tag{17}$$

Proof. For $C_{v,H} = N^{-1}(R^{(v)})^\top R^{(H)}$,

$$\left|\frac{1}{N}\sum_i \langle r_i^{(v)}, r_i^{(H)}\rangle\right| = |\operatorname{tr}(C_{v,H})| \leq \sqrt{d}\|C_{v,H}\|_F.$$

Summing over v and applying Cauchy-Schwarz gives the result. The bound holds for the empirical average over nodes, including dependent node residuals. □

Spectral interpretation. Let $A_Q = \sum_v Q_v Q_v^\top \succeq 0$. The decomposition becomes

$$\mathcal{L}_{\text{orth}} = \operatorname{tr}(h^\top A_Q h) + \frac{1}{M}\operatorname{tr}((h^\top h)^2).$$

For an orthonormal eigensystem (μ_j, u_j) of A_Q, the first term is $\sum_j \mu_j \|u_j^\top h\|_2^2$. The loss therefore penalizes directions with greater relational-disagreement energy more strongly. At fixed Q_v,

$$\nabla_h \mathcal{L}_{\text{orth}} = 2A_Q h + \frac{4}{M}h(h^\top h).$$

During training, the encoder changes both h and the Q_v.

Squared-distance alignment. Expanding $\sum_v \|Q_v - D\|_F^2$ and using $\sum_v Q_v = 0$ gives

$$\sum_v \|Z^{(v)} - Z^{(H)}\|_F^2 = \sum_v \|Q_v\|_F^2 + \frac{(M+1)^2}{M}\|h\|_F^2. \tag{18}$$

For a simple comparison, take $M = 2, N = 2, d = 1$, $Z_R = 0$, $Q_1 = (1, 0)^\top$, and $Q_2 = (-1, 0)^\top$. Choose either $h_a = (1, 0)^\top$ or $h_b = (0, 1)^\top$, and set $Z^{(H)} = 3h/2$. Both configurations have squared-distance alignment 6.5 and $\|h\|_F^2 = 1$, but their orthogonality losses are 2.5 and 0.5. The two penalties share a contraction effect and impose different geometric constraints. InfoNCE also depends on normalization, temperature, and negative examples.

C.2 Conditional perturbation control

Proof of Theorem 1. We first control the change in the subspace selected by P_κ, then add its orthogonal complement. Proposition 1 and the spectral inequality $B_c \succeq \kappa P_\kappa$ give

$$\begin{aligned}\ell_{o,c} &= \frac{1}{N}\operatorname{tr}(h_c^\top B_c h_c) + \frac{1}{MN^2}\|h_c^\top h_c\|_F^2 \\ &\geq \frac{1}{N}\operatorname{tr}(h_c^\top B_c h_c) \geq \kappa\|P_\kappa h_c\|_{\mathrm{rms}}^2.\end{aligned}$$

Combining this control of the corrupted residual with the assumed bound on the clean residual yields

$$\begin{aligned}\|P_\kappa \Delta h\|_{\mathrm{rms}} &\leq \|P_\kappa h_c\|_{\mathrm{rms}} + \|P_\kappa h_0\|_{\mathrm{rms}} \\ &\leq \sqrt{\ell_{o,c}/\kappa} + b_s.\end{aligned}$$

Because P_κ is an orthogonal projector, its selected and complementary components are orthogonal column by column, giving

$$\begin{aligned}\|\Delta h\|_{\mathrm{rms}}^2 &= \|P_\kappa \Delta h\|_{\mathrm{rms}}^2 + \|(I - P_\kappa)\Delta h\|_{\mathrm{rms}}^2 \\ &\leq \left(\sqrt{\ell_{o,c}/\kappa} + b_s\right)^2 + b_n^2.\end{aligned}$$

Taking square roots proves the theorem. When both separation errors vanish, only the loss-controlled term remains. □

The argument applies to changes induced by node attributes, edges, or both, including dependent corruptions. When $P_\kappa = 0$, all perturbation energy lies in the complementary subspace and is bounded by b_n.

The theorem is pointwise for a clean/corrupted pair under a fixed model. A uniform robustness claim would additionally require uniform control of $\ell_{o,c}, b_s, b_n, \kappa$ and the remaining fusion terms over the perturbation set.

C.3 Fusion sensitivity

The residual loss and the final fusion use different stages of the representation. Let $\nu(Z)$ denote row-wise ℓ_2 normalization, with the numerical norm floor used in the implementation. For the same pre-normalization view matrices, define

$$F_g = \frac{\sum_v Z^{(v)} + gZ^{(H)}}{M + g}, \qquad \tilde{Z}_g = \frac{\sum_v \nu(Z^{(v)}) + g\nu(Z^{(H)})}{M + g}.$$

The correction E_g records exactly what this row normalization changes:

$$E_g = \frac{\sum_v [\nu(Z^{(v)}) - Z^{(v)}] + g[\nu(Z^{(H)}) - Z^{(H)}]}{M + g}. \tag{19}$$

For a fixed gate, the anchor drift and normalization-change drift in the corollary are

$$a_c = \|\bar{Z}_c - \bar{Z}_0\|_{\mathrm{rms}}, \qquad e_c = \|E_{g,c} - E_{g,0}\|_{\mathrm{rms}}.$$

Proof of Corollary 1. Writing $D = Z^{(H)} - Z_R$ gives

$$F_g = Z_R + \frac{g}{M+g}D, \qquad \bar{Z} = Z_R + \frac{1}{M+1}D, \qquad h = \frac{M}{M+1}D.$$

Set $\gamma(g) = (1-g)/(M+g)$. Subtracting the anchor from the first identity, and then adding the normalization correction, gives

$$\begin{aligned} F_g &= \bar{Z} - \gamma(g)h, \\ \tilde{Z}_g &= \bar{Z} - \gamma(g)h + E_g. \end{aligned} \tag{20}$$

The trained scalar gate has the same value on clean and corrupted inputs. Subtracting the two fusion identities therefore gives

$$\tilde{Z}_c - \tilde{Z}_0 = (\bar{Z}_c - \bar{Z}_0) - \gamma(g)\Delta h + (E_{g,c} - E_{g,0}).$$

For $g \in [0,1]$, the coefficient $\gamma(g)$ is nonnegative. The three differences are bounded by a_c, $\gamma(g)D_c$, and e_c, respectively, using Theorem 1 for the middle term. The triangle inequality proves the corollary. □

The same calculation also covers an input-dependent gate. In that case, regrouping the residual terms gives

$$\begin{aligned} \tilde{Z}_c - \tilde{Z}_0 = (\bar{Z}_c - \bar{Z}_0) &- \gamma(g_c)\Delta h \\ &- [\gamma(g_c) - \gamma(g_0)]h_0 + (E_{g_c,c} - E_{g_0,0}). \end{aligned}$$

For gate values in $[0,1]$, the triangle inequality gives the more general bound

$$\begin{aligned} \|\tilde{Z}_c - \tilde{Z}_0\|_{\mathrm{rms}} \le a_c + \|E_{g_c,c} - E_{g_0,0}\|_{\mathrm{rms}} + \gamma(g_c)D_c \\ + |\gamma(g_c) - \gamma(g_0)|\|h_0\|_{\mathrm{rms}}. \end{aligned}$$

The extra gate-variation term vanishes when the trained scalar gate is fixed across inputs. For fusion without row normalization, only the anchor-drift and residual-change terms remain. The implemented readout includes row normalization through the correction E_g.

For the pre-normalization average, Proposition 1 also gives

$$\|F_g - \bar{Z}\|_{\mathrm{rms}} \le \gamma(g)(Md\ell_o)^{1/4}.$$

For a fixed prediction map that is L_f-Lipschitz per node, its RMS output change is at most L_f times the representation bound.

C.4 Conditional fusion analysis

A complementary averaging argument concerns errors relative to a latent reference representation, defined through the following statistical model.

Definition 1 (Latent signal-plus-error model)**.** *For a node i, let $z_i^{(k)} = \theta_{*,i} + \epsilon_i^{(k)}$, where $\theta_{*,i}$ is a reference task-relevant representation and the errors have finite second moments. Expectations refer to a specified error distribution for this node.*

Set $G_{k\ell} = \mathbb{E}\langle \epsilon_i^{(k)}, \epsilon_i^{(\ell)} \rangle$. For fixed deterministic weights summing to one, expansion gives

$$\mathrm{MSE}(w) := \mathbb{E}\| \textstyle\sum_k w_k z_i^{(k)} - \theta_{*,i}\|^2 = w^\top G w. \tag{21}$$

The error second-moment matrix G includes contributions from both error variance and bias, so the identity also holds for biased errors.

Assumption 2 (Zero-mean finite-variance errors)**.** $\mathbb{E}\epsilon_i^{(k)} = 0$ *and* $\sigma_k^2 = \mathbb{E}\|\epsilon_i^{(k)}\|^2 < \infty$.

Assumption 3 (Vanishing cross-error inner products)**.** $\mathbb{E}\langle \epsilon_i^{(k)}, \epsilon_i^{(\ell)} \rangle = 0$ *for* $k \neq \ell$.

The following theorem specializes Eq. (21) to zero-mean errors with vanishing cross-error inner products.

Theorem 2 (Fusion with uncorrelated errors)**.** *Under Assumptions 2 and 3, fixed convex weights satisfy*

$$\begin{aligned} \mathrm{MSE}(w) &= \sum_k w_k^2 \sigma_k^2 \le \sigma_{\max}^2, \\ \mathrm{MSE}(\mathbf{1}/K) &= \frac{1}{K^2} \sum_k \sigma_k^2 \le \sigma_{\max}^2 / K, \end{aligned} \tag{22}$$

where $\sigma_{\max}^2 = \max_k \sigma_k^2$*. If every* $\sigma_k^2 > 0$*, the minimizing convex weights and their error are*

$$w_k^* = \frac{1/\sigma_k^2}{\sum_\ell 1/\sigma_\ell^2}, \qquad \mathrm{MSE}(w^*) = \left(\sum_k \frac{1}{\sigma_k^2} \right)^{-1}.$$

Proof. The cross-error assumption removes the off-diagonal terms in Eq. (21), leaving the sum of weighted variances. Convex weights have squared sum at most one, giving the first inequality, and substitution of equal weights gives the second. For positive variances, the objective is strictly convex, and the Lagrange multiplier condition makes $w_k \sigma_k^2$ constant across views, yielding weights proportional to inverse variance. Normalizing gives the stated w_k^*, whose entries are positive and thus satisfy the convex-weight constraint. If a variance is zero, weights supported on zero-variance views attain zero error. □

The equal-weight bound is relative to the largest variance. For example, variances 1 and 100 give averaging error 25.25, compared with best-view error 1.

Theorem 3 (Pairwise distance and fusion error)**.** *Under the same assumptions, let*

$$\mathcal{L}_{\mathrm{pair}} = \binom{K}{2}^{-1} \sum_{k<\ell} \mathbb{E} \| z_i^{(k)} - z_i^{(\ell)} \|^2.$$

Then

$$\mathcal{L}_{\mathrm{pair}} = \frac{2}{K} \sum_k \sigma_k^2, \qquad \mathrm{MSE}(\mathbf{1}/K) = \mathcal{L}_{\mathrm{pair}}/(2K). \tag{23}$$

Proof. For each pair, expected squared distance is $\sigma_k^2 + \sigma_\ell^2$. Each variance occurs in $K-1$ pairs, so division by $K(K-1)/2$ gives the first identity. Substitution into Eq. (22) gives the second. □

Proposition 2 (Cross-error contribution)**.** *Without the zero-mean or vanishing-cross-error assumptions, put* $S = \sum_k G_{kk}$ *and* $C = \sum_{k<\ell} G_{k\ell}$*. Then*

$$\begin{aligned} \mathrm{MSE}(\mathbf{1}/K) &= (S + 2C)/K^2, \\ \mathcal{L}_{\mathrm{pair}} &= 2S/K - 4C/[K(K-1)], \\ \mathrm{MSE}(\mathbf{1}/K) &= \mathcal{L}_{\mathrm{pair}}/(2K) + 2C/[K(K-1)]. \end{aligned} \tag{24}$$

Proof. Set $w = \mathbf{1}/K$ in Eq. (21) for the first identity. Expanding each pairwise distance as $G_{kk} + G_{\ell\ell} - 2G_{k\ell}$, then summing and normalizing, gives the second. Eliminating S gives the third. □

At fixed diagonal second moments, positive C increases the averaging error by $2C/K^2$. For input-dependent random weights, the weights remain inside the expectation.

The quantity $\mathcal{L}_{\mathrm{pair}}$ measures squared Euclidean distances, while the implemented contrastive objective uses projected representations, cosine similarities, and negative examples. The alignment and uniformity perspective (Wang & Isola, 2020) motivates agreement between positive pairs, although transferring the squared-distance identity to the contrastive objective would require additional analysis.

C.5 Residuals and latent errors

With $\bar{\epsilon}_i = K^{-1}\sum_k \epsilon_i^{(k)}$, the observable residual is $r_i^{(k)} = \epsilon_i^{(k)} - \bar{\epsilon}_i$. Expanding the inner product gives the exact relation

$$\mathbb{E}\langle r_i^{(v)}, r_i^{(H)}\rangle = G_{vH} - \frac{1}{K}\sum_k G_{vk} - \frac{1}{K}\sum_k G_{kH} + \frac{1}{K^2}\sum_{k,\ell} G_{k\ell}. \tag{25}$$

For uncorrelated errors with equal second moments σ^2, this is $-\sigma^2/K$. Residual interactions and latent-error cross-moments differ even in this case.

Proposition 3 (Shared-error non-identifiability)**.** *The residual loss alone cannot provide a universal latent-error bound that vanishes as the loss tends to zero.*

Proof. Let every view equal $\Theta_* + B$ for a common nonzero matrix B. All residuals and the loss are zero, while the mean and the pre-normalization gated average retain error B. Without a magnitude restriction, this error can be arbitrarily large. □

The shared-error construction explains why the fusion bound includes an anchor-drift term alongside the contribution controlled by the residual penalty.

D Experimental Settings

D.1 Training protocol

Two-stage training. We follow the standard linear-evaluation protocol for unsupervised multiplex graph learning (Veličković et al., 2019, Shen et al., 2024):

(i) **Pretraining.** The encoder and all MLPs are trained from scratch by minimising Eq. (9) without labels. We use Adam (lr $= 10^{-3}$, weight decay $5 \cdot 10^{-4}$), full-graph batches, for up to 1000 epochs with early stopping (patience 50) on validation $\mathcal{L}_{\text{ssl}}$.

(ii) **Linear probing.** Freeze the encoder and extract $\tilde{Z}$ as in Eq. (4). Train a single linear layer $W \in \mathbb{R}^{d\times C}$ with ℓ_2-regularised softmax cross-entropy on the train split. Hyperparameters are selected on the validation split. Classification metrics in Tables 2 and 3 are reported on the test split.

The controlled experiments use the configuration in Appendix F.1.

D.2 Datasets

We evaluate on four real-world multiplex graphs: ACM and DBLP (Yun et al., 2019), Yelp (Lu et al., 2019), MAG (Wang et al., 2020). Detailed statistics and relational structures are summarised in Table 1. ACM and DBLP are academic heterogeneous networks: ACM uses paper-centered PAP and PSP views, while DBLP uses author-centered APA and APCPA views. Yelp is a standard benchmark for business review analysis. MAG is a large-scale subset of the Microsoft Academic Graph provided by OGB (Hu et al., 2020), focusing on the four most prominent paper categories.

D.3 Hyperparameters

The proposed model is implemented in PyTorch on a workstation with one NVIDIA RTX 4090 (24 GB) GPU and an Intel® Xeon® Platinum CPU. The raw data are transformed into a multi-view configuration: M relational views derived from metapaths and one semantic hypergraph view $\mathcal{H}$ constructed via k-nearest neighbour search ($k = 10$ by default) in feature space. The objective in Eq. (9) comprises the following components: (i) the cross-view contrastive term $\mathcal{L}_{\text{ssl}}^{\text{cross}}$ aligning relational and semantic views. (ii) the augmented-view term $\mathcal{L}_{\text{ssl}}^{\text{aug}}$ encouraging invariance to stochastic feature masking (rate $m = 0.3$). (iii) the residual cross-Gram term $\mathcal{L}_{\text{orth}}$ constraining interactions between relational and semantic residuals. (iv) the predictive consistency term $\mathcal{L}_{\text{cons}}$ with weight $\lambda_c = 0.05$. A Soft Hypergraph Gate (Eq. (4)) is used for fusion. Default values are $\tau = 0.6$, $r = 2$, $d = 512$, $T = 2$ stacking blocks, $\alpha = 0.5$. λ_o is tuned per dataset (searched in $\{0.001, 0.01, 0.1, 1, 10\}$).

D.4 PERTURBATION PROTOCOLS

Edge perturbation (*Edge Ptb*). For a target rate $\eta_{edge} \in \{0.1, 0.2\}$ and each relational view v, we (i) randomly remove $\lceil \eta_{edge} |\mathcal{E}^{(v)}|/2 \rceil$ existing edges and (ii) randomly insert the same number of non-existing edges sampled uniformly from $\mathcal{V} \times \mathcal{V} \setminus \mathcal{E}^{(v)}$. Perturbations are applied to *all* relational views simultaneously. The semantic hypergraph view is rebuilt from the (unperturbed) original features. Five independent perturbation seeds are used per experiment.

Node-feature PGD attack (*Node Ptb*). We follow the standard ℓ_∞-bounded PGD evasion setup (Madry et al., 2018): given a surrogate 2-layer GCN trained on the clean multiplex graph, features X are perturbed to $\tilde{X} = X + \delta$ with $\|\delta\|_\infty \leq \epsilon$, where $\epsilon = \eta_{node} \cdot \mathrm{std}(X)$. PGD runs for $T_{\mathrm{pgd}} = 40$ iterations with step size $\alpha_{\mathrm{pgd}} = \epsilon/10$, using cross-entropy on the validation labels as the attack objective. The resulting $\tilde{X}$ is fed into *all* evaluated models (pretrained on clean graph) at test time. Five random surrogate initialisations are averaged.

Joint-corruption discussion. The edge and feature protocols above are evaluated separately. Joint corruption can affect both views in correlated ways, since relational encoders and the KNN semantic view share node features. An exact-component experiment examines simultaneous structural and feature perturbations under a separate training protocol, as detailed in Appendix F.

D.5 COMPUTATIONAL COST

Direct residual cross-Gram products cost $\mathcal{O}(MNd^2)$ and require $d \times d$ cross-Gram matrices. Predictive consistency uses two masked-embedding paths through the auxiliary head. Measured ACM runtime and memory for the controlled configuration appear in Appendix F.

E ADDITIONAL RESULTS

E.1 CLUSTERING

We also evaluate the frozen embeddings $\tilde{Z}$ (Eq. (4)) using K-means clustering and report ACC, NMI, F1, and ARI. We compare against unsupervised baselines: (1) *Fundamental*: VGAE (Kipf & Welling, 2016), DGI (Veličković et al., 2019). (2) *Multi-view*: O2MAC (Fan et al., 2020), MvAGC (Lin & Kang, 2021), MCGC (Pan & Kang, 2021), HDMI (Jing et al., 2021). (3) *Disentangled/GSL*: MGDCR (Mo et al., 2023a), DMG (Mo et al., 2023b), BTGF (Qian et al., 2024), InfoMGF (Shen et al., 2024).

Table 6 reports clustering results. OCG leads all four metrics on ACM, DBLP, and MAG. On Yelp, InfoMGF leads NMI, ACC, and F1, while OCG leads ARI.

E.2 RELATED REGULARIZERS

Barlow Twins (Zbontar et al., 2021) encourages the cross-correlation matrix between augmented representations to approach the identity, combining diagonal invariance and off-diagonal redundancy reduction. VICReg (Bardes et al., 2022) combines invariance with variance and covariance terms. CCA-style methods (Andrew et al., 2013) seek correlated representations. Ortho-GConv (Guo et al., 2022) constrains feature transformations within GNN layers. OCG penalizes every entry of the cross-Gram matrix between relational and semantic residuals centered at a multi-view mean, constraining the geometry of relative view differences that can include both useful information and noise.

Appendix F.2 compares the two terms in the OCG residual objective.

E.3 VIEW AGREEMENT AND COLLAPSE

Zero residuals arise whenever all views share the same representation, whether that representation varies across nodes or is constant. InfoNCE supplies a discrimination signal through positive and negative node pairs, with a constant representation giving loss $\log N$ for one N-candidate term. The residual penalty is scale-sensitive, motivating the normalized interaction diagnostic in Eq. (26).

Table 6: Node clustering results (%). The top three values in each column use boldface, underline, and color, respectively. OOM denotes out of memory under the evaluated configuration.

Method	ACM				DBLP			
	NMI	ARI	ACC	F1	NMI	ARI	ACC	F1
VGAE	45.83	41.36	67.93	68.62	61.79	65.56	84.48	83.67
DGI	52.94	47.55	65.36	57.34	65.59	70.35	86.88	86.02
O2MAC	42.36	46.04	77.92	78.01	58.64	60.01	83.29	82.88
MvAGC	64.49	66.81	87.17	87.21	50.39	51.21	78.39	77.84
MCGC	60.21	50.72	65.62	54.78	65.56	71.51	87.96	87.47
HDMI	65.44	68.87	88.11	88.14	64.85	70.85	87.39	86.75
MGDCR	58.80	55.15	73.82	70.34	62.47	62.22	81.91	80.16
DMG	64.14	67.21	87.11	87.23	69.03	73.07	88.45	87.88
BTGF	68.92	73.14	90.09	90.11	66.28	72.47	88.05	87.28
InfoMGF	74.89	81.09	92.82	92.89	70.19	73.49	88.72	88.31
OCG	**77.06**	**81.94**	**93.59**	**93.64**	**70.78**	**73.61**	**89.76**	**89.35**

Method	Yelp				MAG			
	NMI	ARI	ACC	F1	NMI	ARI	ACC	F1
VGAE	39.19	42.57	65.07	56.74	OOM	OOM	OOM	OOM
DGI	39.42	42.62	65.29	56.79	53.56	42.60	59.89	57.17
O2MAC	39.02	42.53	65.07	56.74	OOM	OOM	OOM	OOM
MvAGC	24.39	29.25	63.14	56.70	OOM	OOM	OOM	OOM
MCGC	38.35	35.17	65.61	57.49	OOM	OOM	OOM	OOM
HDMI	60.81	59.35	79.56	77.60	48.15	34.92	51.78	49.80
MGDCR	44.23	46.47	72.71	54.43	54.43	43.98	61.37	60.53
DMG	65.66	66.33	88.26	89.27	48.72	39.77	61.61	60.16
BTGF	69.97	73.53	91.39	92.32	OOM	OOM	OOM	OOM
InfoMGF	**72.67**	74.66	**91.85**	**92.86**	56.65	45.25	64.13	63.09
OCG	69.98	**75.83**	91.46	91.49	**57.32**	**46.11**	**68.65**	**68.26**

F MECHANISM AND EFFICIENCY

F.1 PROTOCOL AND DIAGNOSTICS

The controlled ACM experiments examine residual geometry using two-layer GCN encoders with hidden dimension 128, representation dimension 64, and dropout 0.5. Training lasts 200 epochs with initial learning rate 0.01, cosine scheduling, and no weight decay. The semantic view combines feature-neighborhood hypergraphs with $k = 3$ and $k = 5$, weighted by 0.6 and 0.4, respectively. The mechanism experiment disables refinement ($T = 0$) and assigns coefficient 3×10^{-9} to the residual penalty or its selected exact component, holding the remaining configuration fixed across variants.

The consistency paths independently mask the view embeddings, normalize them, and fuse them with equal weights before a linear head with dropout. The head uses softmax temperature 0.6, and consistency is enabled after epoch 100 with weight 0.05. The contrastive semantic weights and final linear-probe readout still use the learned scalar gate.

We train the encoder without label-dependent gradient updates, then evaluate the frozen embeddings with a linear probe. The probe uses learning rate 0.1, no weight decay, at most 1000 epochs, patience 100, and a fixed initialization seed of 12345. Encoder checkpoints are evaluated every 10 epochs and selected by validation classification accuracy, so validation labels are used for model selection. We report test-set results for training seeds $0, 1, 2$, using their arithmetic mean and sample standard deviation (denominator $n - 1$). For joint corruption, the corruption seed varies with the training seed, so the reported dispersion reflects variation across matched training and corruption seeds.

Table 7: Exact-component ablation on ACM with joint corruption and support overlap $\rho = 1$. Mean $\pm$ sample standard deviation over three seeds. C_{res} is normalized residual cross-Gram interaction. Residual magnitude is $K^{-1}\sum_k \|R^{(k)}\|_F/\sqrt{N}$.

Penalty	Macro-F1 (%)	C_{res}	Residual magnitude
None	90.15 ± 0.36	0.423 ± 0.011	0.689 ± 0.017
Original	90.34 ± 0.16	0.369 ± 0.017	0.626 ± 0.013
Coupling only	90.57 ± 0.51	0.421 ± 0.022	0.683 ± 0.010
Spectral only	90.31 ± 1.04	0.380 ± 0.033	0.629 ± 0.006

Table 8: Micro-F1 (%) for the exact-component experiment in Table 7. Three-seed mean $\pm$ sample standard deviation.

Penalty	None	Original	Coupling only	Spectral only
Micro-F1	90.06 ± 0.35	90.23 ± 0.18	90.46 ± 0.51	90.20 ± 1.06

For residuals centered at the empirical view mean, we measure

$$C_{\text{res}} = \frac{1}{M}\sum_{v=1}^{M} \frac{\|(R^{(v)})^\top R^{(H)}\|_F}{\max\{\|R^{(v)}\|_F \|R^{(H)}\|_F, \varepsilon\}}, \qquad A_{\text{res}} = \frac{1}{K}\sum_{k=1}^{K} \frac{\|R^{(k)}\|_F}{\sqrt{N}}, \tag{26}$$

where ε is a numerical safeguard. C_{res} measures the magnitude of normalized residual cross-Gram interactions. A_{res} averages the root-mean-square node-vector norm across views, using normalization by $\sqrt{N}$.

F.2 Exact-component experiment

For the experiment in Table 7, all features of 10% of ACM nodes are masked to zero. Structural perturbations are restricted to edges incident to a selected set of the same size. Each relational view has separate deletion and insertion budgets, each set to 10% of its original undirected edge count. Deletions are sampled among eligible incident edges. Insertion candidates come from the union of original relational edges that are absent after deletion, restricted to the selected nodes. Deleted edges can be reinserted, so the final number of distinct edge changes may be smaller than the operation budget. The semantic view is constructed from the resulting features, and the model is trained on the corrupted inputs.

Let S_X denote the feature-masked nodes and S_E the nodes eligible for incident-edge perturbation. With equal set sizes, the support-overlap parameter is $\rho = |S_X \cap S_E|/|S_X|$, up to rounding. Table 7 fixes $\rho = 1$ and compares twelve runs, using three seeds for each of the four configurations. Table 8 supplies the corresponding Micro-F1 values. The spectral term is $M^{-1}\|(R^{(H)})^\top R^{(H)}\|_F^2$.

Table 7 compares the original objective with its coupling-only and spectral-only components at the same coefficient. The spectral term reduces residual interaction and magnitude to values close to those of the original penalty. Coupling-only achieves the highest mean classification score with smaller changes in the residual diagnostics. Differences in classification scores remain small relative to the variability across three seeds. The diagnostic results suggest that spectral shrinkage accounts for much of the change in residual geometry.

F.3 Efficiency

Table 9 reports measurements on clean ACM with an NVIDIA RTX 3090, using two refinement blocks for OCG. After five warm-up steps, training time is averaged over thirty steps of the training loss and inference time over thirty embedding-extraction calls. FLOPs estimate the main graph-learning, encoding, and refinement operations, with one multiply and one addition counted as two

operations. Top-k selection, elementwise operations, and residual diagnostics are excluded. Memory and parameter storage use 2^{20} bytes per MiB.

Table 9: Computational cost on clean ACM. Peak memory is the measured GPU allocation. Timings are within-run averages.

Configuration	Parameters	Size (MiB)	Inference (GFLOPs)	Train step (ms)	Inference (ms)	Peak GPU (MiB)
InfoMGF	320,780	1.22	79.69	143.02	12.64	2148.0
OCG w/o $\mathcal{L}_{\mathrm{orth}}$	382,792	1.46	119.83	229.09	19.51	3174.6
OCG	382,792	1.46	119.83	229.01	17.83	3174.6

The residual penalty adds no parameters, and both profiled OCG variants use the same embedding-extraction routine, including residual-loss diagnostics. The timings compare repeated executions of the same inference procedure, and the complete OCG configuration incurs additional cost relative to InfoMGF. An inference-only ACM subgraph sweep with $N = 1000, 2000, 3025$ gives $14.17, 53.47, 119.83$ GFLOPs and $6.34, 11.03, 17.97$ ms, respectively. Scaling beyond this ACM range requires measurements on larger graphs.

G LIMITATIONS

The residual bound depends on directional separation and the loss evaluated on corrupted inputs, quantities that are not estimated by the experiments. Shared errors, corruption outside the penalized directions, and useful information within them can limit the benefit of regularization. Anchor drift and normalization effects require separate control in the fusion bound. The robustness evaluation covers random edge changes and transfer feature attacks. The controlled mechanism study uses a smaller model without refinement, limiting direct comparison with the benchmark configuration.

The auxiliary consistency loss admits constant-output heads, so its contribution is measured through ablation alongside the contrastive objectives that supply the representation-learning signal.