FORGE AND BRIDGE: STRICTLY INDUCTIVE MULTIMODAL INTENT DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised multimodal intent discovery aims to uncover potential intent categories from unlabeled multimodal interactions and generalize the discovered structure to unseen samples. A key challenge lies in the fact that although multimodal large language models can provide rich sample-level semantic descriptions, if we directly rely on MLLMs for classification, it often leads to category collapse. In this paper, we propose a unified framework named IntentForge–IntentBridge, which separates semantic descriptions, intent category discovery, and prediction for unseen samples: IntentForge organizes samples based on their consensus and organizes them into a structured semantic view, from which fixed training partitions are discovered, while IntentBridge migrates this partition to unseen samples through reliability-guided pointwise prediction. The framework uses MLLMs to describe what each sample means, thereby mitigating dominant-category attraction, redundant categories, and fragmented intent structures caused by direct MLLM categorization, and discovers the intent taxonomy from semantic consensus across the entire training set. Our algorithm achieves 62.92%, 47.70%, and 39.49% accuracy on MIntRec, MELD-DA, and IEMOCAP-DA. Especially on the MIntRec dataset, the accuracy rate is far superior to that of existing works, and it also achieves excellent performance on the MELD-DA dataset.

1 INTRODUCTION

Human communicative intent depends on words, tone, visual behavior, and context (Baltrusaitis et al., 2019; Zhang et al., 2022). The same words can convey either praise or sarcasm in different contexts (Castro et al., 2019), while dialogue history helps interpret an utterance (Ghosal et al., 2019; Lu et al., 2020). For example, "You're really impressive" may express genuine praise when spoken sincerely, but could convey sarcasm when delivered with an exaggerated or dismissive tone. Conversely, different phrasings can also serve the same communicative purpose. Tensor and low-rank fusion model cross-modal interactions (Zadeh et al., 2017; Liu et al., 2018); cross-modal attention handles unaligned sequences (Tsai et al., 2019), and representation disentanglement and information maximization address complementary modality content (Hazarika et al., 2020; Han et al., 2021). These supervised advances motivate richer utterance representations, but discovering intent categories without annotations additionally requires deciding which communicative meanings belong together.

Recent discovery methods improve multimodal representations: UMC combines multimodal augmentation with reliable sample selection (Zhang et al., 2024a), and GLAD aligns global and local semantics (Zhu et al., 2026). Their features leave communicative purpose implicit. Meanwhile, CLIP and ALIGN align visual and textual semantics (Radford et al., 2021; Jia et al., 2021), and ImageBind extends a shared embedding space across modalities (Girdhar et al., 2023). BLIP and BLIP-2 connect visual understanding with language generation (Li et al., 2022; 2023); Flamingo and LLaVA further support multimodal prompting and instruction following (Alayrac et al., 2022; Liu et al., 2023). These capabilities make explicit utterance descriptions feasible, but do not themselves establish a coherent intent taxonomy. MCSP introduces explicit semantic guidance by generating concepts from cluster representatives with an MLLM and propagating this guidance to refine representations (Gu et al., 2026). However, its concept generation depends on clusters and representatives derived from the initial feature space. Intent distinctions poorly captured in this space may therefore be underrepresented in the evidence presented to the MLLM. This motivates generating semantic descriptions directly from each utterance's multimodal content before clustering.

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.

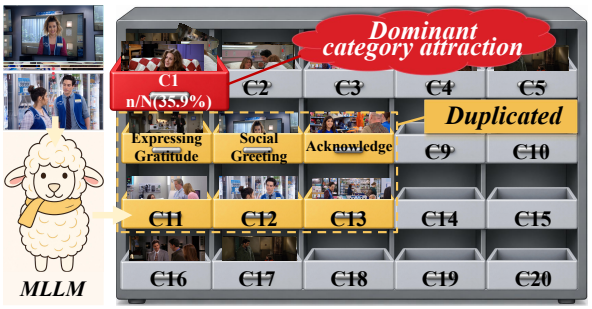


Figure 1: Per-sample plausibility does not guarantee dataset-level taxonomy consistency. Direct MLLM categorization produces dominant-category attraction, semantically duplicated classes, and fragmented tail categories.

Table 1: Indicative MIntRec results (%). Direct-MLLM and IntentForge–IntentBridge illustrate sequential category generation and frozen pointwise assignment, respectively. Best results are shown in red, and second-best results are underlined.

Method	ACC ↑	NMI ↑	ARI ↑
MiniCPM-o	40.09	48.15	16.26
Gemma-4	<u>51.69</u>	<u>57.22</u>	<u>32.29</u>
Qwen3-VL	36.83	46.53	17.22
Phi-4-MM	29.16	30.19	8.66
Ours	**62.92**	**63.16**	**41.63**

Can an MLLM instead assign categories directly? In our MIntRec diagnostic, we first ask four MLLMs to classify utterances directly into 20 categories, but they produce fewer categories than requested. Enforcing exactly 20 categories leads to taxonomy collapse, with redundant categories and highly imbalanced assignments. For example, greetings are split across six categories, while C1 contains most samples (Figure 1; Table 1; Appendix F). These observations motivate separating semantic description from category formation: MLLMs describe individual utterances, and subsequent clustering determines the dataset-level partition. This design grounds category formation in sample-level descriptions while reducing reliance on MLLMs' autonomous category decisions.

We separate description from assignment. A frozen MLLM produces a structured *Card* for each utterance, recording an open-ended description of observable characteristics and source-grounded evidence without predefined target category names or provisional memberships. Shared fields make utterances comparable while retaining their specific meanings. **IntentForge–IntentBridge** (IntentFB) then discovers categories from agreement among Card-derived views and transcripts on training data. IntentBridge learns to transfer this fixed partition through reliability-guided student–EMA training and exemplar retrieval. All artifacts are frozen before inference, and each unseen utterance is assigned independently.

Our main contributions are as follows:

- We organize each user's statement into a Card, construct high-quality semantic representations, and then perform clustering, effectively addressing category collapse.
- We develop IntentForge–IntentBridge to discover training categories from Card-based semantic consensus and transfer the fixed partition to unseen utterances.
- We achieved excellent results on three benchmark datasets and show how IntentForge improves category discovery and IntentBridge improves predictions for unseen samples.

2 Related Work

Multimodal intent discovery requires both extracting clear intent semantics from heterogeneous inputs and organizing them into coherent categories that generalize to unseen samples, motivating our review of large-model-assisted semantic construction and multi-view clustering with partition transfer. **Large-model-assisted multimodal clustering.** Large language and multimodal models provide semantic knowledge that can help clustering move beyond feature similarity (Ke et al., 2024). In text clustering, Viswanathan et al. (2024) investigate language-model guidance through feature enrichment, clustering constraints, and assignment correction. ClusterLLM uses triplet judgments to improve embeddings and pairwise judgments to determine clustering granularity (Zhang et al., 2023). Related work extends semantic guidance to images: TAC retrieves textual knowledge from WordNet and exchanges neighborhood information between text and image features (Li et al., 2024), while Chu et al. (2024) combine pretrained multimodal features with rate-reduction clustering and semantic naming. For multimodal intent discovery, MCSP selects representatives from initial clusters, generates concepts through MLLM reasoning, and propagates these concepts to refine representations (Gu

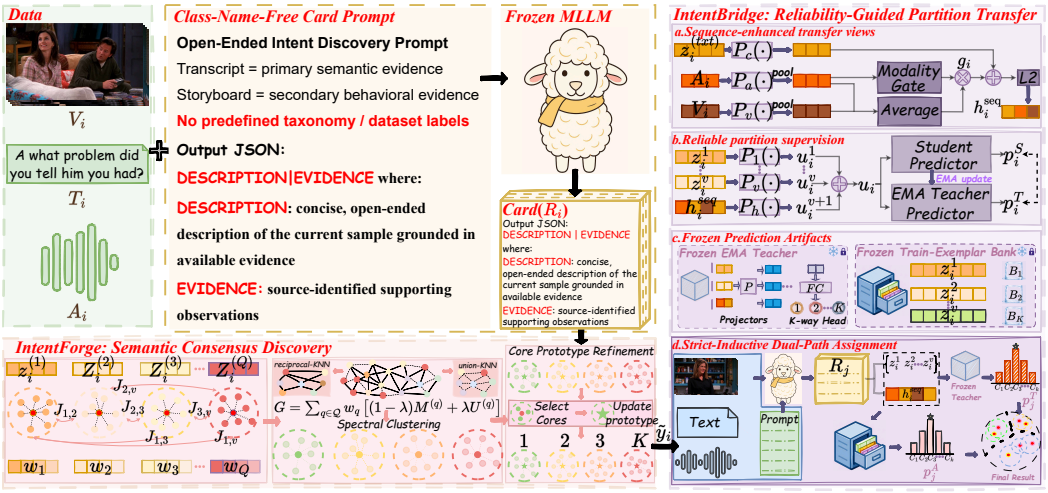


Figure 2: IntentForge–IntentBridge. IntentForge forms fixed anonymous assignments $\tilde{\mathbf{y}}$ by consensus clustering, core-based prototype refinement, and train-only propagation. IntentBridge learns from curriculum anchors and freezes an EMA predictor and exemplar banks. Each query is assigned using predictor and retrieval distributions $\mathbf{p}_j^T, \mathbf{p}_j^A$, without test–test relations.

et al., 2026). Although these methods enrich clustering with semantic knowledge, deriving guidance from provisional cluster representatives can miss intent distinctions obscured by the initial grouping, motivating independent utterance-level descriptions before category formation. **Multi-view clustering and partition transfer.** Multi-view clustering exploits complementary information through co-regularization, cross-modal supervision, and joint representation learning (Kumar et al., 2011; Alwassel et al., 2020; Chen et al., 2021; Luo et al., 2024; Ren et al., 2025), with extensions for missing views and semantic consistency (Lin et al., 2021; Xu et al., 2022; Pu et al., 2024). For multimodal intent discovery, UMC combines modality-masked augmentation with density-based sample selection to guide representation learning (Zhang et al., 2024a). GLAD captures shared semantic structure through global alignment and weights local relationships by sample importance (Zhu et al., 2026). Related approaches connect representations with cluster assignments through joint optimization and centroid guidance (Zhang et al., 2021; 2024b), while Mean Teacher uses weight-averaged models to provide consistency targets (Tarvainen & Valpola, 2017). Although these methods improve representations and cluster formation through cross-view agreement, they do not explicitly address how to reliably transfer a fixed discovered partition to unseen samples processed independently. Existing approaches leave a gap between utterance-level semantic evidence and reliable assignment of unseen samples, which IntentForge–IntentBridge addresses through Card-based consensus discovery and reliability-guided partition transfer.

3 METHOD

3.1 PROBLEM SETUP

In unlabeled $\mathcal{D}_{\text{tr}} = \{x_i\}_{i=1}^N$, each $x_i = (T_i, \mathbf{A}_i, \mathbf{V}_i, u_i)$ contains text, audio, video, and prescribed context. We assume known $K \geq 2$, without predefined target category names or instance labels; estimating K is outside scope. Our goal is to divide the multi-view data according to the communicative intent into $K \geq 2$ categories, grouping samples with the same communicative intent into one category, and learning classification rules applicable to unseen samples. For each new sample, the model independently predicts its classification in these categories based on its multimodal information.

Semantic Card construction. Before category discovery, a frozen MLLM converts each utterance into a structured semantic Card:

$$R_i = M_{\text{frozen}}(\Xi(x_i); P_{\text{card}}) \,. \tag{1}$$

Here, $\Xi(x_i)$ inserts the current transcript into the shared template and attaches the available media from the current sample. P_{card} is the same unsupervised prompt across all three datasets and requests

an open-ended description of observable characteristics with supporting evidence. The prompt, schema, and illustrative examples appear in Appendix E.

3.2 INTENTFORGE: SEMANTIC CONSENSUS DISCOVERY

Card-based semantic views. Given the Cards $\{R_i\}_{i=1}^N$, fixed constructors $s_i^{(q)} = \psi_q(R_i, T_i)$ form complementary views from `description`, the original transcript, and visual observations in `evidence`. Quoted transcript spans in `evidence` support grounding checks. The active view set $\mathcal{Q}$ excludes unavailable modalities before training and remains fixed at inference. A frozen text encoder E gives

$$\mathbf{z}_i^{(q)} = \mathrm{Norm}\Big(E(s_i^{(q)})\Big), \qquad q \in \mathcal{Q}. \tag{2}$$

Norm is row-wise ℓ_2-normalization. Constructors and evidence checks are in Appendices C and E.

Neighborhood consensus. Let $\mathcal{N}_{i,h}^{(q)}$ be the h highest-cosine training neighbors of sample i in view q, excluding itself. View agreement is

$$J_{qr}^{(h)} = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|\mathcal{N}_{i,h}^{(q)} \cap \mathcal{N}_{i,h}^{(r)}\right|}{\left|\mathcal{N}_{i,h}^{(q)} \cup \mathcal{N}_{i,h}^{(r)}\right|}. \tag{3}$$

For $Q = |\mathcal{Q}| \geq 2$, define $u_q = [\max((Q-1)^{-1}\sum_{r\neq q} J_{qr}^{(k_\mathrm{w})}, \epsilon_\mathrm{w})]^{1/2}$ and $\bar{w}_q = (1 - Q\delta_\mathrm{w})\,\mathrm{softmax}_q(\log \mathbf{u}/\tau_\mathrm{w}) + \delta_\mathrm{w}$, where $\mathbf{u} = (u_q)_{q\in\mathcal{Q}}$. Reserve $w_\mathrm{txt} = \gamma_\mathrm{txt}$ and set $w_q = (1-\gamma_\mathrm{txt})\bar{w}_q / \sum_{r\neq\mathrm{txt}} \bar{w}_r$ for non-text views. This reserves a fixed text reference; with only two active semantic views, the other mass is also fixed at $1-\gamma_\mathrm{txt}$. Agreement still determines graph mixing. All weighting parameters are preset (Appendix D.1).

Construct directed affinities $D_{ij}^{(q)} = \mathbb{1}[j \in \mathcal{N}_{i,k_\mathrm{g}}^{(q)}][\mathrm{clip}((1 + \mathbf{z}_i^{(q)\top}\mathbf{z}_j^{(q)})/2, 0, 1)]^2$, and element-wise reciprocal and union graphs $M^{(q)} = \min(D^{(q)}, D^{(q)\top})$, $U^{(q)} = \max(D^{(q)}, D^{(q)\top})$. With $\bar{J}^{(k_\mathrm{g})}$ the mean agreement over distinct view pairs, $\lambda = \mathrm{clip}((\bar{J}^{(k_\mathrm{g})} - a_\lambda)/b_\lambda, 0, 1)$ interpolates from reciprocal to union edges as agreement increases:

$$\mathbf{G} = \sum_{q\in\mathcal{Q}} w_q \left[(1-\lambda)\mathbf{M}^{(q)} + \lambda \mathbf{U}^{(q)}\right], \qquad G_{ii} = 1. \tag{4}$$

The discovery representation uses the same view weights:

$$\mathbf{f}_i = \mathrm{Norm}\Big(P_\mathrm{disc}\Big[\mathrm{concat}_{q\in\mathcal{Q}}\,\sqrt{w_q}\,\mathbf{z}_i^{(q)}\Big]\Big). \tag{5}$$

P_disc is training-mean-centered PCA to d_disc, fitted only to weighted training concatenations.

Fixed partition construction. Spectral clustering (Ng et al., 2001) of $\mathbf{G}$ yields K groups. Within each group, rank members by cosine similarity to its normalized mean and average the selected central members into a normalized prototype. Reassign samples to their nearest prototype while restoring selected members to their original group, until convergence or T_ref iterations. Appendix D.1 specifies selection ratios and size caps. From the refined labels $\widehat{y}_i^{\mathrm{ref}}$, select central anchors $\mathcal{C}_k$ and initialize $S_i^{(0)} = \mathbf{e}_{\widehat{y}_i^{\mathrm{ref}}}$ for anchors and zero otherwise, where $\mathbf{e}_k$ is the k-th one-hot vector. In discovery space, let $\mathcal{N}_{i,k_\mathrm{prop}}$ contain the k_prop nearest training neighbors excluding i, and set $W_{\mathrm{prop},ij} = \mathbb{1}[j \in \mathcal{N}_{i,k_\mathrm{prop}}]\exp[-(1-\mathbf{f}_i^\top\mathbf{f}_j)/\tau_\mathrm{prop}]$. Using neighborhood label propagation (Zhou et al., 2003), update

$$\mathbf{S}^{(t+1)} = \mathrm{RowNorm}\Big(\mathbf{W}_\mathrm{prop}^\top \mathbf{S}^{(t)}\Big). \tag{6}$$

where RowNorm normalizes nonzero rows and leaves zero rows unchanged. After each update, reset anchors to their initial one-hot labels. The fixed targets are $\widetilde{y}_i = \arg\max_k S_{ik}^{(T_{\mathrm{prop}})}$. Proceed only if all samples form exactly K nonempty groups; tie-breaking and boundary cases are in Appendix D.2.

Theorem 1 (Conditional recovery by anchored propagation)**.** *Fix a reference partition $y_i^\star \in [K]$. Suppose anchors have correct labels, every edge $W_{\mathrm{prop},uv} > 0$ connects equal reference labels, and each node is reachable from an anchor along these edges. If D is the maximum shortest anchor-to-node distance, Equation 6 with anchor resets recovers the reference partition for $T_{\mathrm{prop}} \geq D$.*

This theorem shows how reliable anchors and within-class connectivity jointly enable label propagation to recover the full partition when every node is reachable from an anchor. Proof and failure cases appear in Appendix B.1.

3.3 IntentBridge: Reliability-Guided Partition Transfer

IntentForge supervises IntentBridge with fixed anonymous training assignments $\widetilde{\mathbf{y}}$. IntentBridge retains this partition and learns a pointwise predictor through reliability-guided student–EMA training.

Sequence-enhanced transfer views. To incorporate acoustic and visual sequence features into transfer, project the transcript to $\mathbf{c}_i = P_c(\mathbf{z}_i^{(\mathrm{txt})})$ and encode acoustic and visual sequences with attention pooling, with $\mathbf{c}_i, \mathbf{a}_i, \mathbf{v}_i \in \mathbb{R}^{d_{\mathrm{seq}}}$. With a learned scalar network F_g, fixed bias b_g, and $g_i = \sigma(F_g([\mathbf{c}_i; \mathbf{a}_i; \mathbf{v}_i]) - b_g)$, fuse the unnormalized projected vectors as

$$\mathbf{h}_i^{(\mathrm{seq})} = \mathrm{Norm}\left(\mathbf{c}_i + g_i \frac{\mathbf{a}_i + \mathbf{v}_i}{2}\right). \tag{7}$$

Let C, A, V stack normalized $\mathbf{c}_i, \mathbf{a}_i, \mathbf{v}_i$. The branch minimizes symmetric cross-modal contrastive losses ℓ_τ and a text-reference residual penalty $\mathcal{L}_{\mathrm{res}}$:

$$\mathcal{L}_{\mathrm{align}} = \ell_{\tau_{\mathrm{align}}}(C, A) + \ell_{\tau_{\mathrm{align}}}(C, V) + \lambda_{\mathrm{av}} \ell_{\tau_{\mathrm{align}}}(A, V) + \lambda_{\mathrm{res}} \mathcal{L}_{\mathrm{res}}. \tag{8}$$

Both losses are defined in Appendix D.1. The sequence encoders, transcript projection, pooling, and gate are then frozen. Appendix B.2 bounds residual deviation under a norm condition. For $\mathcal{V} = \mathcal{Q} \cup \{\mathrm{seq}\}$, set $\mathbf{x}_i^{(q)} = \mathbf{z}_i^{(q)}$ for semantic views and $\mathbf{x}_i^{(\mathrm{seq})} = \mathbf{h}_i^{(\mathrm{seq})}$. Preset nonnegative masses ν_q, normalized over available views, give

$$\boldsymbol{\phi}_i = \mathrm{Norm}\Big(P_{\mathrm{trans}}\Big[\mathrm{concat}_{q \in \mathcal{V}} \sqrt{\nu_q}\, \mathbf{x}_i^{(q)}\Big]\Big). \tag{9}$$

P_{trans} is a separate train-fitted, centered PCA to dimension d_{trans}. Add the shared view $\mathbf{x}_i^{(\mathrm{sh})} = \boldsymbol{\phi}_i$ to obtain $\mathcal{U} = \mathcal{V} \cup \{\mathrm{sh}\}$. The preset ν_q are distinct from discovery weights w_q.

Reliable partition supervision. For $\mathcal{I}_k = \{i : \widetilde{y}_i = k\}$, score local same-group density, group cohesion, separation from competing groups, cross-view label support, and stability across preset clustering runs. Let $\mathcal{S} = \{\mathrm{dens}, \mathrm{coh}, \mathrm{margin}, \mathrm{view}, \mathrm{stab}\}$ index these five deterministic training-set scores $r_{i,s} \in [0, 1]$, defined in Appendix C; stability aligns to fixed anonymous IDs, never gold labels. With $\chi_i \in \{0, 1\}$ indicating that the record was not rejected by the checks in Appendix E, reliability is their validity-gated geometric mean:

$$r_i = \chi_i \left[\prod_{s \in \mathcal{S}} \max(r_{i,s}, \epsilon_{\mathrm{r}})\right]^{1/|\mathcal{S}|}, \tag{10}$$

Here $\epsilon_{\mathrm{r}} > 0$ is a numerical floor. At epoch e, select the highest-reliability members with $\chi_i = 1$ in each group into $\mathcal{A}_k^{(e)}$, expanding the selected fraction linearly from η_0 to $\eta_{\max}$ over E_{cur} epochs. Exact counts and schedules are in Appendix D.1; the partition itself never changes.

Student–EMA learning. Learn projectors P_q into a common prediction space and fusion logits $\boldsymbol{\ell} \in \mathbb{R}^{|\mathcal{U}|}$; define global view weights $\gamma_q = \mathrm{softmax}(\boldsymbol{\ell})_q$:

$$\mathbf{g}_i^S = \sum_{q \in \mathcal{U}} \gamma_q P_q(\mathbf{x}_i^{(q)}). \tag{11}$$

The student outputs $\mathbf{h}_i^S = \mathrm{Norm}(\mathbf{g}_i^S)$ and $\mathbf{p}_i^S = \mathrm{softmax}(W\mathbf{g}_i^S + \mathbf{b})$, with learned classifier $W, \mathbf{b}$. Following Mean Teacher (Tarvainen & Valpola, 2017), initialize the teacher from the student and update all its projectors, gate logits, and classifier after each step by $\theta_T \leftarrow m_{\mathrm{EMA}}\theta_T + (1 - m_{\mathrm{EMA}})\theta_S$. Here $m_{\mathrm{EMA}} \in [0, 1)$; stop-gradient teacher targets are computed at the start of each epoch and held fixed within that epoch. For fixed targets $\mathbf{q}_i = \mathbf{e}_{\widetilde{y}_i}$, define group support $\bar{r}_k = |\mathcal{A}_k^{(e)}|^{-1} \sum_{i \in \mathcal{A}_k^{(e)}} r_i$, $\rho_k = \sqrt{\max(\bar{r}_k, \epsilon_\omega)} / \sum_l \sqrt{\max(\bar{r}_l, \epsilon_\omega)}$, and within-group weights $\omega_i = \rho_{\widetilde{y}_i} r_i / \max(\sum_{j \in \mathcal{A}_{\widetilde{y}_i}^{(e)}} r_j, \epsilon_\omega)$, where $\epsilon_\omega > 0$ guards denominators. The partition loss is

$$\mathcal{L}_P = \sum_{k=1}^{K} \sum_{i \in \mathcal{A}_k^{(e)}} \omega_i \, \mathrm{KL}(\mathbf{q}_i \| \mathbf{p}_i^S). \tag{12}$$

The full objective adds teacher-target KL loss $\mathcal{L}_T$, feature-dropout cosine consistency $\mathcal{L}_{\mathrm{cons}}$, and cluster-prior matching $\mathcal{L}_{\mathrm{bal}}$:

$$\mathcal{L} = \mathcal{L}_P + \lambda_T(e)\mathcal{L}_T + \lambda_C \mathcal{L}_{\mathrm{cons}} + \lambda_B \mathcal{L}_{\mathrm{bal}}. \tag{13}$$

Appendix D.1 defines auxiliary losses and teacher-weight warm-up. Train for E_{train} epochs without partition refresh or early stopping, then freeze the final EMA predictor. Per group, store up to $B_{\max}$ highest-reliability members of $\mathcal{A}_k^{(E_{\mathrm{train}})}$, with views, reliability, and IDs, in a nonempty bank $\mathcal{B}_k$.

3.4 Strictly Inductive Inference

Process query j independently with the frozen feature pipeline and EMA predictor to obtain $\{\mathbf{x}_j^{(q)}\}_{q \in \mathcal{U}}$ and $\mathbf{p}_j^T$. For each view and group, retrieve the $\min(L, |\mathcal{B}_k|)$ highest-cosine bank members $\mathcal{T}_{jk}^{(q)}$. Their reliability-weighted score is

$$s_{jk}^{(q)} = \frac{\sum_{a \in \mathcal{T}_{jk}^{(q)}} r_a \cos(\mathbf{x}_j^{(q)}, \mathbf{x}_a^{(q)})}{\max(\sum_{a \in \mathcal{T}_{jk}^{(q)}} r_a, \epsilon_A)}. \tag{14}$$

With $\epsilon_A > 0$ and temperature $\tau_A > 0$, set $\mathbf{p}_j^{(q)} = \mathrm{softmax}(\mathbf{s}_j^{(q)} / \tau_A)$, where $\mathbf{s}_j^{(q)} = (s_{jk}^{(q)})_{k=1}^K$. Using the transfer masses ν_q and preset shared-view mass $\lambda_{\mathrm{sh}} \in [0, 1]$, aggregate retrieval as

$$\mathbf{p}_j^A = (1 - \lambda_{\mathrm{sh}}) \sum_{q \in \mathcal{V}} \nu_q \mathbf{p}_j^{(q)} + \lambda_{\mathrm{sh}} \mathbf{p}_j^{(\mathrm{sh})}. \tag{15}$$

Let $c(\mathbf{p}) = 1 - H(\mathbf{p}) / \log K$, where H is Shannon entropy, and set $\beta_j = c(\mathbf{p}_j^T) / \max(c(\mathbf{p}_j^T) + c(\mathbf{p}_j^A), \epsilon_{\mathrm{conf}})$, with $\epsilon_{\mathrm{conf}} > 0$. The query-specific fusion gives

$$\widehat{y}_j = \arg\max_{k \in [K]} [\beta_j \mathbf{p}_j^T + (1 - \beta_j)\mathbf{p}_j^A]_k. \tag{16}$$

If both confidences vanish, $\beta_j = 0$; ties use the smallest ID. Other boundary cases: Appendix D.2.

Theorem 2 (Margin preservation under dual-path fusion)**.** *Let $\mathbf{p}^F = \beta\mathbf{p}^T + (1 - \beta)\mathbf{p}^A$ for the query's fusion weight, and let y uniquely maximize $\mathbf{p}^T$. For $m_T = p_y^T - \max_{k \neq y} p_k^T > 0$ and $\Delta = \|\mathbf{p}^A - \mathbf{p}^T\|_\infty$,*

$$p_y^F - \max_{k \neq y} p_k^F \geq m_T - 2(1-\beta)\Delta. \tag{17}$$

Thus $m_T > 2(1-\beta)\Delta$ preserves the teacher's prediction.

This auxiliary stability result preserves the teacher's prediction, not its correctness, and does not imply an accuracy gain from fusion. Proof and the case of agreeing paths are in Appendix B.3.

Pointwise information flow. Let Ω_{tr} collect the frozen Card prompt and generator, encoders, train-fitted transforms, view settings, EMA parameters, IDs, and training banks. Equation 16 is $F(x_j; \Omega_{\mathrm{tr}})$, with no test–test relations or test-time updates. Under deterministic, evaluation-mode pointwise operations and fixed query context, predictions are invariant to other test samples (Appendix B.4); this does not guarantee correctness or replace an implementation audit.

4 Experiments

4.1 Datasets and experimental configuration.

We conducted the evaluation on three representative multimodal semantic understanding benchmarks: MIntRec (Zhang et al., 2022) is a multimodal intent recognition dataset containing 20 fine-grained intents, while MELD-DA (Poria et al., 2019) and IEMOCAP-DA (Busso et al., 2008) are constructed based on MELD and IEMOCAP respectively, both containing 12 dialogue-act categories. All three cover three modalities: text, audio, and vision. Following the settings of existing unsupervised multimodal semantic discovery work, we adopted a training/test split ratio of approximately 4:1. The numbers of training/test samples for MIntRec, MELD-DA, and IEMOCAP-DA are 1,779/445, 7,990/1,998, and 7,532/1,884 respectively, and the corresponding numbers of clustering categories K are 20, 12, and 12. Unless otherwise specified, all ablation experiments, task comparisons, and other experiments adopt the same training set division, preprocessing, evaluation protocol, and applicable settings as the baseline comparison. UMC, and IntentFB are run five times under this unified benchmark configuration, with seed values of $\{0, 1, 2, 3, 4\}$; since GLAD and MCSP do not have their codes publicly available, their published results are retained. For IntentFB, frozen MiniCPM-o-4.5 is used to generate cards (Cards), and frozen BGE-M3 is used to encode the cards, followed by classification. Shared hyperparameters are in Table 6, while input budgets and generation settings are detailed in Appendices D and E. We report mean $\pm$ standard deviation over five matched seeds, using Hungarian-aligned ACC and ARI, NMI, and FMI for clustering evaluation.

4.2 Baseline comparison.

Table 2 compares IntentFB with SCCL (Zhang et al., 2021), CC (Li et al., 2021), USNID (Zhang et al., 2024b), MCN (Chen et al., 2021), UMC (Zhang et al., 2024a), GLAD (Zhu et al., 2026), and MCSP (Gu et al., 2026); Appendix G records result sources. IntentFB reaches 62.92/47.70/39.49% ACC on MIntRec/MELD-DA/IEMOCAP-DA. It leads all four metrics on MIntRec, exceeding GLAD by 16.95 ACC points, and leads ACC, NMI, and FMI on MELD-DA. On IEMOCAP-DA, its higher ACC accompanies lower ARI, NMI, and FMI than GLAD. Although the effect on IEMOCAP-DA is average, it is mainly because the information of IEMOCAP-DA is too fragmented, making it difficult for the model to summarize useful information from the fragmented data.

4.3 Controlled ablation and IntentBridge comparison.

Tables 3 and 4 respectively evaluated the performance of IntentForge and IntentBridge under conditions of shared cards, semantic features, context, input budget, data partitioning, preprocessing, K value, and random seed. Table 3 shows that, under the nearest-centroid assignment allocation method, after introducing IntentForge, the accuracy rates of MIntRec and MELD-DA increased by 11.28 and 14.10 points respectively, indicating that IntentForge significantly improved the quality of training partition discovery. With the addition of IntentBridge, the accuracy rates further increased by 1.80 and 1.51 points, achieving the best overall result in the 2×2 comparison. Table 4 further indicates that when the partitions of IntentForge are fixed, IntentBridge achieves the highest average ACC, ARI, and NMI on both datasets, and outperforms the strongest simple allocation scheme by 1.75 points on

Table 2: Native-system benchmark results. Superscripts distinguish model variants below. Bold/underline denote the highest/second-highest displayed values; dashes denote unavailable results.

Method	MIntRec				MELD-DA				IEMOCAP-DA			
	ACC	ARI	NMI	FMI	ACC	ARI	NMI	FMI	ACC	ARI	NMI	FMI
SCCL (2021)	36.63	16.25	44.20	23.80	29.50	13.38	21.72	23.98	26.89	11.01	22.09	24.56
CC (2021)	40.67	21.20	47.37	26.21	24.99	13.04	23.03	24.37	25.31	12.46	22.87	23.92
USNID (2024b)	40.36	20.97	47.51	26.37	24.39	12.16	20.65	24.88	27.17	11.48	22.03	23.59
MCN (2021)	16.90	1.95	18.85	9.81	18.18	1.56	8.49	15.41	15.87	1.73	8.01	13.99
UMC (2024a)	43.37	24.35	48.67	28.69	34.14	20.50	21.40	32.51	35.49	21.35	24.49	34.40
MCSP[1] (2026)	44.54	24.85	49.07	29.61	34.62	22.07	21.49	33.74	–	–	–	–
MCSP[2] (2026)	45.21	25.59	48.42	30.28	34.57	21.57	21.61	33.16	–	–	–	–
GLAD (2026)	<u>45.97</u>	<u>26.59</u>	<u>52.17</u>	<u>31.84</u>	<u>37.79</u>	**24.39**	<u>25.63</u>	<u>36.16</u>	<u>38.26</u>	**24.84**	**28.22**	**37.37**
IntentFB (ours)	**62.92**	**41.63**	**63.16**	**45.43**	**47.70**	<u>24.09</u>	**29.25**	**40.32**	**39.49**	20.11	<u>24.55</u>	<u>37.05</u>

MCSP[1]/MCSP[2]: Qwen3-VL/Gemini.

MIntRec and 1.51 points on MELD-DA. This supports its role in more effectively transferring the discovered partitions to unseen samples. Overall, these controlled experimental results support the complementary role of the two modules: IntentForge provides a dominant advantage by constructing stronger training partitions, while IntentBridge continuously improves the point-to-point partition migration performance, achieving the best overall performance under shared semantic input.

Table 3: Controlled 2×2 ablation with shared Cards, semantic features, context, and input budgets. Forge/Bridge denote IntentForge/IntentBridge; values are mean $\pm$ standard deviation over five matched seeds. Bold/underline mark the highest/second-highest means.

	Components		MIntRec			MELD-DA		
	Forge	Bridge	ACC↑	ARI↑	NMI↑	ACC↑	ARI↑	NMI↑
Item-1	×	×	49.84±1.3	29.15±1.8	53.39±1.2	32.09±1.1	15.60±0.6	24.93±0.7
Item-2	√	×	<u>61.12±0.5</u>	<u>39.65±0.9</u>	<u>61.61±1.1</u>	<u>46.19±1.5</u>	<u>23.44±0.8</u>	<u>28.90±0.6</u>
Item-3	×	√	50.88±1.2	30.09±1.4	54.14±1.7	36.59±1.7	17.74±1.7	26.51±0.6
Item-4	√	√	**62.92±1.0**	**41.63±1.2**	**63.16±1.5**	**47.70±1.1**	**24.09±1.3**	**29.25±0.9**

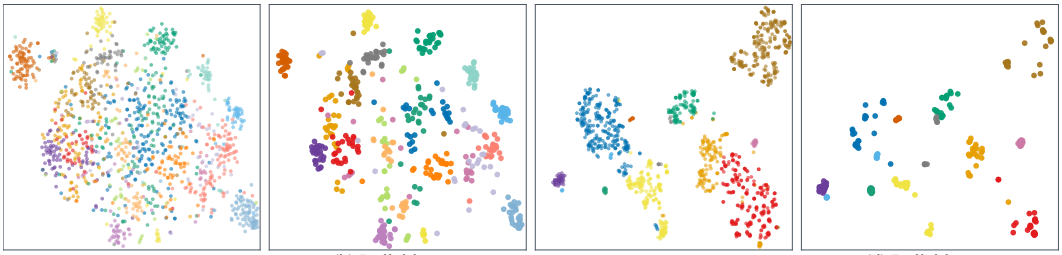
Table 4: Assignment comparison with shared inputs and a fixed IntentForge partition per seed. $\mathcal{N}$: nearest centroid; $\mathcal{L}$: linear classifier; $\mathcal{R}$: retrieval only; $\mathcal{B}$: full IntentBridge.

	MIntRec			MELD-DA		
Method	ACC↑	ARI↑	NMI↑	ACC↑	ARI↑	NMI↑
$\mathcal{N}$	61.12±0.5	39.65±0.9	61.61±1.1	46.19±1.5	23.44±0.8	28.90±0.6
$\mathcal{L}$	61.17±0.4	33.30±1.4	59.56±0.7	44.57±0.2	21.06±0.2	26.26±0.2
$\mathcal{R}$	59.69±0.6	37.93±0.7	59.31±0.6	45.45±0.1	21.77±0.1	27.96±0.1
$\mathcal{B}$	**62.92±1.0**	**41.63±1.2**	**63.16±1.5**	**47.70±1.1**	**24.09±1.3**	**29.25±0.9**

4.4 Qualitative Analysis and Retrieval Sensitivity

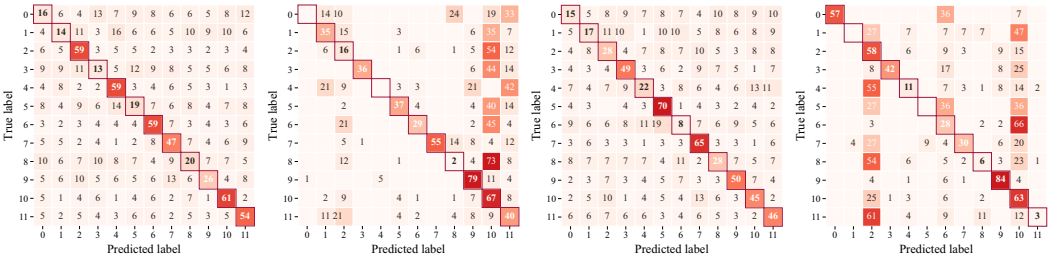
Figure 3 shows that the highest-reliability cores form more compact clusters than the full exemplar distributions on MIntRec and IEMOCAP-DA. This demonstrates that our reliability-guided selection identifies representative samples for reliable partition transfer and retrieval-based inference. Figure 4 compares IntentFB with GLAD on MELD-DA and IEMOCAP-DA, showing ACC improvements from 37.79% to 47.70% and from 38.26% to 39.49%, respectively. The value of GLAD is derived from its original text. These results demonstrate the effectiveness of our framework in improving overall classification accuracy, although macro recall decreases on both datasets. Figure 5 evaluates the effects of retrieval limit L and temperature τ_A on ACC across three datasets, showing competitive

performance under multiple parameter configurations. This indicates that our retrieval mechanism can maintain effective classification performance without relying on a single parameter combination. Overall, these results indicate that IntentFB can select representative samples, improve classification accuracy, and achieve significant results under different parameter configurations.



(a) All exemplars MIntRec (b) Reliable cores MIntRec (c) All exemplars IEMOCAP-DA (d) Reliable cores IEMOCAP-DA

Figure 3: Frozen exemplars and highest-reliability cores on shared t-SNE coordinates for MIntRec (a,b) and IEMOCAP-DA (c,d), coloured by anonymous cluster.



(a) GLAD (MELD-DA) ACC = 37.79% | mR = 37.25%
(b) Ours (MELD-DA) ACC = **47.70%** | mR = 32.91%
(c) GLAD (IEMOCAP-DA) ACC = 38.26% | mR = 36.92%
(d) Ours (IEMOCAP-DA) ACC = **39.49%** | mR = 31.69%

Figure 4: Row-normalized confusion matrices (%) after post-hoc Hungarian alignment: GLAD versus IntentFB on MELD-DA and IEMOCAP-DA. Red boxes mark diagonals.

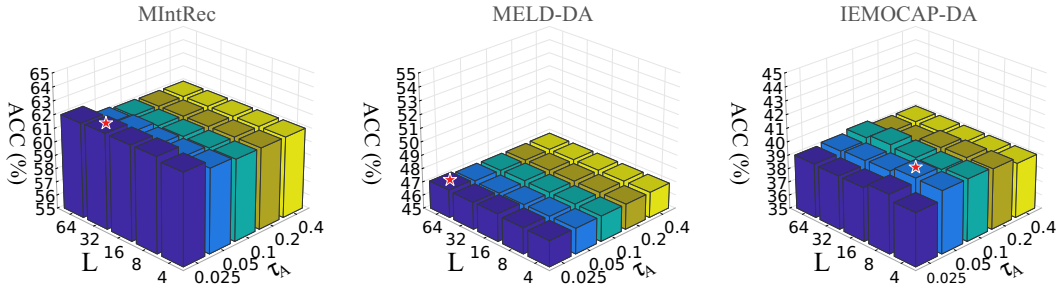


Figure 5: Joint sensitivity to retrieval limit L and temperature τ_A. Bars show ACC on MIntRec, MELD-DA, and IEMOCAP-DA; stars from the source plot do not denote statistical significance.

5 Conclusion

We propose IntentFB, a strictly inductive framework for unsupervised multimodal intent discovery that separates utterance-level semantic description, training-category discovery, and pointwise assignment of unseen samples. Extensive comparisons, controlled ablations, and mechanism analyses demonstrate its effectiveness, with category partitioning providing the main gains and category transfer offering further improvements. However, MLLM-generated Cards cannot fully prevent category collapse, as semantically similar fine-grained intents may remain difficult to distinguish and concentrate in a few categories. Future work will combine Card-based semantic guidance with UMC-style contrastive learning, using MLLM-generated Cards as teacher signals to better separate closely related intents and further alleviate category collapse.

AI USE STATEMENT

We used generative AI tools to assist with experimental infrastructure, including generating scripts to run our implementation across multiple datasets and supporting the execution of comparative experiments. The proposed method and its implementation were developed and written by the authors. The manuscript and theoretical proofs were also written by the authors. AI assistance in manuscript preparation was limited to identifying formatting and spelling errors and providing feedback on the correctness of our author-written proofs. AI feedback was used as an aid to checking, not as a substitute for mathematical justification. The authors take responsibility for the final manuscript, theoretical claims, implementation, and reported experimental results.As a component of the proposed framework, we also used frozen MiniCPM-o-4.5 to generate semantic Cards from individual samples, as described in Section 4.1 and Appendices D and E. The generation configuration, prompt, and validation procedures are documented there.

REPRODUCIBILITY STATEMENT

We thank the reviewers for taking the time to read our appendices. They provide theoretical proofs (Appendix B), a summary of notation (Appendix C), implementation details and algorithmic pseudocode (Appendices D and A), the Card-generation prompt and illustrative examples (Appendix E), direct-MLLM diagnostics (Appendix F), comparison protocols and ablation studies (Appendix G), additional reproducibility details, local-generator diagnostics (Appendix H), and mechanism experiments (Appendix I).

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*, pp. 23716–23736, 2022.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-Supervised Learning by Cross-Modal Audio-Video Clustering. In *NeurIPS*, pp. 9758–9770, 2020.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *TPAMI*, pp. 423–443, 2019.

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. EDA: Enriching emotional dialogue acts using an ensemble of neural annotators. In *LREC*, pp. 620–627, 2020.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *LRE*, pp. 335–359, 2008.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards Multimodal Sarcasm Detection (An *Obviously* Perfect Paper). In *ACL*, pp. 4619–4629, 2019.

Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, James Glass, Michael Picheny, and Shih-Fu Chang. Multimodal Clustering Networks for Self-Supervised Learning From Unlabeled Videos. In *ICCV*, pp. 8012–8021, 2021.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of ACL*, pp. 2318–2335, 2024.

Tianzhe Chu, Shengbang Tong, Tianjiao Ding, Xili Dai, Benjamin Haeffele, Rene Vidal, and Yi Ma. Image Clustering via the Principle of Rate Reduction in the Age of Pretrained Models. In *ICLR*, pp. 24140–24162, 2024.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *EMNLP-IJCNLP*, pp. 154–164, 2019.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All. In *CVPR*, pp. 15180–15190, 2023.

Yunjin Gu, Qianrui Zhou, and Hua Xu. Unsupervised Multimodal Intent Discovery via MLLM-Guided Concept Generation and Semantic Propagation. *arXiv*, 2026.

Wei Han, Hui Chen, and Soujanya Poria. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *EMNLP*, pp. 9180–9192, 2021.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *ACM MM*, pp. 1122–1131, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML*, pp. 4904–4916, 2021.

JunLong Ke, Zichen Wen, Yechenhao Yang, Chenhang Cui, Yazhou Ren, Xiaorong Pu, and Lifang He. Integrating Vision-Language Semantic Graphs in Multi-View Clustering. In *IJCAI*, pp. 4273–4281, 2024.

Abhishek Kumar, Piyush Rai, and Hal Daume III. Co-regularized Multi-view Spectral Clustering. In *NeurIPS*, pp. 1413–1421, 2011.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, pp. 12888–12900, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, pp. 19730–19742, 2023.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive Clustering. In *AAAI*, pp. 8547–8555, 2021.

Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image Clustering with External Guidance. In *ICML*, pp. 27890–27902, 2024.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. COMPLETER: Incomplete Multi-View Clustering via Contrastive Prediction. In *CVPR*, pp. 11174–11183, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *NeurIPS*, pp. 34892–34916, 2023.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *ACL*, pp. 2247–2256, 2018.

Junyu Lu, Xiancong Ren, Yazhou Ren, Ao Liu, and Zenglin Xu. Improving Contextual Language Models for Response Retrieval in Multi-Turn Conversation. In *SIGIR*, pp. 1805–1808, 2020.

Caixuan Luo, Jie Xu, Yazhou Ren, Junbo Ma, and Xiaofeng Zhu. Simple Contrastive Multi-View Clustering with Data-Level Fusion. In *IJCAI*, pp. 4697–4705, 2024.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On Spectral Clustering: Analysis and an Algorithm. In *NeurIPS*, pp. 849–856, 2001.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *ACL*, pp. 527–536, 2019.

Jingyu Pu, Chenhang Cui, Xinyue Chen, Yazhou Ren, Xiaorong Pu, Zhifeng Hao, Philip S. Yu, and Lifang He. Adaptive Feature Imputation with Latent Graph for Deep Incomplete Multi-View Clustering. In *AAAI*, pp. 14633–14641, 2024.

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*, pp. 8748–8763, 2021.

Yazhou Ren, Junlong Ke, Zichen Wen, Tianyi Wu, Yang Yang, Xiaorong Pu, and Lifang He. Multi-View Graph Clustering via Node-Guided Contrastive Encoding. In *ICML*, pp. 51421–51435, 2025.

Antti Tarvainen and Harri Valpola. Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. In *NeurIPS*, pp. 1195–1204, 2017.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *ACL*, pp. 6558–6569, 2019.

Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. Large Language Models Enable Few-Shot Clustering. *TACL*, pp. 321–333, 2024.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-Level Feature Learning for Contrastive Multi-View Clustering. In *CVPR*, pp. 16051–16060, 2022.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor Fusion Network for Multimodal Sentiment Analysis. In *EMNLP*, pp. 1103–1114, 2017.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. Supporting Clustering with Contrastive Learning. In *NAACL*, pp. 5419–5430, 2021.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. MIntRec: A New Dataset for Multimodal Intent Recognition. In *ACM MM*, pp. 1688–1697, 2022.

Hanlei Zhang, Hua Xu, Fei Long, Xin Wang, and Kai Gao. Unsupervised Multimodal Clustering for Semantics Discovery in Multimodal Utterances. In *ACL*, pp. 18–35, 2024a.

Hanlei Zhang, Hua Xu, Xin Wang, Fei Long, and Kai Gao. A Clustering Framework for Unsupervised and Semi-Supervised New Intent Discovery. *TKDE*, pp. 5468–5481, 2024b.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. ClusterLLM: Large Language Models as a Guide for Text Clustering. In *EMNLP*, pp. 13903–13920, 2023.

Dengyong Zhou, Olivier Bousquet, Thomas N. Lal, Jason Weston, and Bernhard Schölkopf. Learning with Local and Global Consistency. In *NeurIPS*, pp. 321–328, 2003.

Zhengzhong Zhu, Pei Zhou, Weihong Du, Shiquan Min, and Jiangping Zhu. Unsupervised Semantic Discovery via Global and Local Semantic Alignment in Multimodal Clustering. In *AAAI*, pp. 35248–35256, 2026.

We thank the reviewers for taking the time to read our appendices. They present theoretical proofs (Appendix B), notation (Appendix C), implementation details and pseudocode (Appendices D and A), the Card prompt and examples (Appendix E), direct-MLLM diagnostics (Appendix F), comparison protocols and ablations (Appendix G), local-generator diagnostics (Appendix H), and mechanism experiments (Appendix I).

A Algorithmic Overview

Algorithm 1 summarizes the complete IntentForge–IntentBridge pipeline in Section 3. IntentForge first discovers a fixed anonymous partition from Card-derived multi-view representations through consensus clustering, prototype refinement, and anchored propagation (Lines 1–7). IntentBridge then learns and freezes reliability-guided transfer artifacts via sequence-enhanced views, student–EMA training, and exemplar-bank construction (Lines 8–20). At inference, all learned artifacts remain frozen, and each unseen query is assigned independently by combining teacher prediction with group-wise retrieval (Lines 21–26), preserving the strictly inductive setting without test–test interaction. All fitting and reliability calculations use training data only.

Algorithm 1 IntentForge–IntentBridge

Input: Unlabeled training set $\mathcal{D}_{\mathrm{tr}}$, group count K, frozen generator and semantic encoder, preset configuration.
Output: Frozen artifacts Ω_{tr} and anonymous query IDs $\widehat{y}_j$.

1: Generate Cards and fixed available views with Equation 1 and Equation 2; record validity χ_i.
2: Compute neighbor agreement (Equation 3) and weights w_q, reserving the text mass.
3: Form G (Equation 4) and PCA features $\mathbf{f}_i$ (Equation 5); spectrally cluster G into K groups.
4: Refine central-member prototypes; reassign and restore selected members to source groups until convergence or T_{ref} iterations.
5: Select propagation anchors $\mathcal{C}_k$; initialize their one-hot scores and zero non-anchor scores; build directed W_{prop}.
6: Apply Equation 6 for T_{prop} steps, resetting anchors after each update.
7: Fix $\widetilde{y}_i = \arg\max_k S_{ik}^{(T_{\mathrm{prop}})}$, using smallest-ID ties; require all samples and K nonempty groups.
8: Train and freeze the sequence branch using Equation 7 and Equation 8.
9: Fit P_{trans}, obtain $\boldsymbol{\phi}_i$ with Equation 9, and add the shared view to $\mathcal{U}$.
10: Compute five reliability signals and r_i (Equation 10); apply Appendix D.2 checks.
11: Initialize the student (Equation 11) and EMA teacher $\theta_T \leftarrow \theta_S$.
12: **for** epoch $e = 1, \ldots, E_{\mathrm{train}}$ **do**
13: Select valid, reliable $\mathcal{A}_k^{(e)}$ by expanding quotas; compute ρ_k, ω_i (Equation 12).
14: Cache epoch-start stop-gradient teacher targets (Appendix D.1).
15: **for** minibatch $\mathcal{M}$ in one shuffled training pass **do**
16: Optimize θ_S with Equation 13 and the minibatch scaling in Appendix D.2.
17: Update $\theta_T \leftarrow m_{\mathrm{EMA}}\theta_T + (1 - m_{\mathrm{EMA}})\theta_S$.
18: **end for**
19: **end for**
20: Build nonempty $\mathcal{B}_k$ from the top $B_{\max}$ available members of final $\mathcal{A}_k$ by reliability, storing views, r_i, and IDs; freeze banks, final teacher, and all Ω_{tr}.
21: **for each** unseen query x_j, independently, **do**
22: Obtain query views $\{\mathbf{x}_j^{(q)}\}_{q\in\mathcal{U}}$ and $\mathbf{p}_j^T$ with the frozen pipeline.
23: Retrieve $\min(L, |\mathcal{B}_k|)$ nearest bank members per view and group; score with Equation 14.
24: Apply retrieval softmax and aggregate $\mathbf{p}_j^A$ with Equation 15.
25: Compute entropy-confidence β_j; return $\widehat{y}_j$ by Equation 16, using smallest-ID ties.
26: **end for**

B Theoretical Analysis and Proofs

The results in Section 3 are auxiliary, conditional analyses of individual operations, not a general recovery or performance guarantee for the complete method. They separate conditional partition recovery, representation preservation, prediction stability, and test-cohort invariance. None assumes that unlabeled agreement scores are calibrated probabilities of semantic correctness. In particular, no result establishes unconditional recovery of human intent labels or improved generalization from EMA training. The residual theorem applies to the text-anchored branch; archived variants that use a different fusion rule require a separate analysis.

B.1 Proof of Theorem 1

Proof. Let d_i be the shortest directed anchor-to-node distance and $D = \max_i d_i$. Write $W = W_{\text{prop}}$ and let $\mathcal{C}$ be the union of the clamped anchor sets. An edge $u \to v$ means $W_{uv} > 0$, which is the direction in which information flows because the update uses $W^\top S$. At initialization, each anchor row equals $\mathbf{e}_{y_i^\star}$, and each non-anchor row is zero.

First, induction on t shows that every row of $S^{(t)}$ belongs to $\{\mathbf{0}, \mathbf{e}_{y_i^\star}\}$. This holds at $t = 0$. For a non-anchor i, its unnormalized next row is

$$(W^\top S^{(t)})_i = \sum_u W_{ui} S_u^{(t)} = \left(\sum_{u : S_u^{(t)} \neq \mathbf{0}} W_{ui} \right) \mathbf{e}_{y_i^\star}. \tag{18}$$

The second equality follows from edge purity: each positive summand has $y_u^\star = y_i^\star$. Row normalization yields either the same one-hot vector or zero. Anchor resetting also preserves the claimed property.

Next, induction on shortest directed distance establishes eventual permanent activation. Anchors, whose distance is zero, are clamped at every iteration. For a non-anchor at distance $d \geq 1$, a shortest path provides a predecessor u at distance $d - 1$ with $W_{ui} > 0$. By the distance induction hypothesis, $S_u^{(t)}$ is its correct one-hot vector for every $t \geq d - 1$. Thus the scalar in Equation 18 is strictly positive for every such t, and $S_i^{(t+1)} = \mathbf{e}_{y_i^\star}$. Taking the largest distance proves recovery for $T_{\text{prop}} \geq D$. □

Scope and failure cases. The theorem concerns the propagation kernel W_{prop}, not the different consensus graph G. It does not assert that spectral clustering or core-based prototype refinement supplies correct anchors. A wrongly labeled anchor remains wrong because it is clamped. Without directed anchor reachability, a node may retain a zero row and receive the smallest ID through tie-breaking. Cross-group edges can inject competing label mass, invalidating the one-hot induction. The graph's direction is essential: reachability must follow $u \to v$ when $W_{uv} > 0$, rather than the reverse direction.

B.2 Residual Deviation

Theorem 3 (Bounded deviation of the sequence residual)**.** *For a fixed utterance, write* $\mathbf{d}_i^{\text{res}} = g_i(\mathbf{a}_i + \mathbf{v}_i)/2$ *and assume* $\mathbf{c}_i \neq \mathbf{0}$ *and* $\alpha_i = \|\mathbf{d}_i^{\text{res}}\|_2 / \|\mathbf{c}_i\|_2 < 1$*. When normalization uses the ordinary Euclidean norm, the residual in Equation 7 satisfies*

$$1 - \cos(\mathbf{c}_i, \mathbf{h}_i^{(\text{seq})}) \leq 1 - \sqrt{1 - \alpha_i^2}, \qquad \left\| \mathbf{h}_i^{(\text{seq})} - \frac{\mathbf{c}_i}{\|\mathbf{c}_i\|_2} \right\|_2 \leq \sqrt{2(1 - \sqrt{1 - \alpha_i^2})}. \tag{19}$$

Proof. Suppress the sample index and set $\mathbf{u} = \mathbf{c}/\|\mathbf{c}\|_2$ and $\mathbf{e} = \mathbf{d}^{\text{res}}/\|\mathbf{c}\|_2$, so $\|\mathbf{u}\|_2 = 1$ and $\|\mathbf{e}\|_2 = \alpha < 1$. Then $\|\mathbf{u} + \mathbf{e}\|_2 \geq 1 - \alpha > 0$, and $\mathbf{u}^\top(\mathbf{u} + \mathbf{e}) \geq 1 - \alpha > 0$. Consequently the angle θ between $\mathbf{u}$ and $\mathbf{h} = (\mathbf{u} + \mathbf{e})/\|\mathbf{u} + \mathbf{e}\|_2$ lies in $[0, \pi/2)$.

The distance from $\mathbf{u}$ to the line spanned by the unit vector $\mathbf{h}$ equals $\sqrt{1 - (\mathbf{u}^\top \mathbf{h})^2} = \sin\theta$. Since $\mathbf{u} + \mathbf{e}$ lies on that line, this distance is no larger than $\|\mathbf{e}\|_2 = \alpha$. Positivity of $\cos\theta$ gives $\cos\theta \geq \sqrt{1 - \alpha^2}$. The first claim follows, and the identity

$$\|\mathbf{h} - \mathbf{u}\|_2^2 = 2 - 2\mathbf{u}^\top \mathbf{h} = 2(1 - \cos\theta) \tag{20}$$

gives the second claim. □

Connection to the gate and loss. The triangle inequality gives the directly checkable sufficient bound

$$\alpha_i \le \frac{g_i(\|\mathbf{a}_i\|_2 + \|\mathbf{v}_i\|_2)}{2\|\mathbf{c}_i\|_2}. \tag{21}$$

For any batch satisfying the theorem's conditions, the mean of $1 - \sqrt{1-\alpha_i^2}$ bounds $\mathcal{L}_{\text{res}}$ from above. Separately, with unit normalization the mean squared displacement from the normalized transcript reference is exactly $2\mathcal{L}_{\text{res}}$. These relations explain what the regularizer measures.

Protected normalization. The Euclidean-distance bound assumes unit output vectors. If implementation uses $v/\max(\|v\|_2, \varepsilon)$, it applies when the relevant norms exceed the guard. A sufficient condition for the residual norm is $\|\mathbf{c}_i\|_2(1-\alpha_i) > \varepsilon$. Zero or guard-dominated vectors are outside this unit-normalization result.

B.3 Proof of Theorem 2

Proof. For a probability vector on $K \ge 2$ classes, $0 \le H(\mathbf{p}) \le \log K$. Hence the confidences are nonnegative, and the guarded definition of β gives $0 \le \beta \le 1$. Write $\mathbf{d} = \mathbf{p}^A - \mathbf{p}^T$. For every competitor $k \ne y$,

$$\begin{aligned} p_y^F - p_k^F &= p_y^T - p_k^T + (1-\beta)(d_y - d_k) && (22)\\ &\ge m_T - (1-\beta)(|d_y| + |d_k|) && (23)\\ &\ge m_T - 2(1-\beta)\Delta. && (24) \end{aligned}$$

Taking the minimum over competitors proves Equation 17. A strictly positive lower bound makes y the unique fused maximizer. If both paths instead share a unique maximizer y, then each pairwise margin is positive in both paths, and

$$p_y^F - p_k^F = \beta(p_y^T - p_k^T) + (1-\beta)(p_y^A - p_k^A) > 0 \tag{25}$$

for every $\beta \in [0,1]$, including the endpoints. □

Symmetry and interpretation. Interchanging the paths gives the retrieval-preservation condition $m_A > 2\beta\Delta$, where m_A is the retrieval path's unique-winner margin. Both criteria hold for the actual query-derived weight; the proof does not assume that β is independent of either distribution. They certify stability for a given pair of outputs, not across parameter changes that may also alter both outputs and β.

B.4 Test-Cohort Invariance

Theorem 4 (Test-cohort invariance)**.** *Condition on fixed training artifacts Ω_{tr}. Assume inference is deterministic, uses evaluation-mode pointwise operations, and keeps each query's prescribed evidence fixed independently of its test cohort. For any two ordered test collections containing the same query x, Equation 16 assigns x the same anonymous ID. Predictions are therefore equivariant to test permutations and consistent between any retained subset and the full collection.*

Proof. Fix Ω_{tr} and a query x, including its prescribed context. The evidence organizer, fixed prompt, and deterministic frozen generator return the same Card for x in either test collection. The string constructors and frozen encoders therefore return the same semantic views. Evaluation-mode sequence modules, train-fitted PCA maps, and pointwise normalization return the same remaining representations.

The frozen EMA predictor consequently returns the same $\mathbf{p}^T$. Retrieval compares these fixed query views only with fixed training-bank members. With fixed tie-breaking, the retrieved sets, their reliability weights, the scores in Equation 14, and the aggregate $\mathbf{p}^A$ are identical. Entropy confidence and the final smallest-ID tie rule are deterministic functions of these two probability vectors. Thus the assigned ID is the same function $F(x; \Omega_{\text{tr}})$ in both collections.

For an ordered collection $X = (x_1, \dots, x_n)$, define $\mathcal{F}(X) = (F(x_1; \Omega_{\text{tr}}), \dots, F(x_n; \Omega_{\text{tr}}))$. Applying the pointwise identity to each entry gives $\mathcal{F}(\sigma X) = \sigma\mathcal{F}(X)$ for any permutation σ, and $\mathcal{F}(X|_I) = \mathcal{F}(X)|_I$ for any retained index set I. □

Implementation boundary. The statement excludes test-batch normalization statistics, stochastic generation driven by a cohort-dependent random stream, test-bank updates, and context constructed from whichever test queries happen to be present. If inference is stochastic, fixed randomness attached to each query gives the corresponding conditional statement; uncontrolled randomness does not give exact repeated-call equality. Floating-point and batching effects must still be checked in an implementation audit. The theorem concerns anonymous predictions before scoring; a post-hoc Hungarian mapping may change when the evaluated subset changes.

C Notation and Operational Definitions

Table 5 lists key symbols; Appendix D.1 defines reliability and Table 6 records core settings. All dataset-level statistics and learned transformations use training data only; the resulting inference state is frozen before evaluation. Norm denotes row-wise Euclidean normalization with a protected norm.

Table 5: Key notation. Computational definitions are given in Appendix D.1.

Symbol	Brief definition
$\mathcal{Q}, \mathcal{V}, \mathcal{U}$	Semantic views; semantic plus sequence views; those views plus the shared representation, respectively.
R_i, ψ_q	Semantic Card and three view constructors: description, current transcript, and visual evidence (Appendix E).
$\mathbf{z}_i^{(q)}, \mathbf{f}_i$	Frozen semantic embedding and discovery representation, respectively.
$\mathbf{h}_i^{(\mathrm{seq})}, \boldsymbol{\phi}_i$	Aligned sequence representation and shared transfer representation, respectively.
$\widetilde{y}_i, \mathcal{I}_k$	Fixed anonymous assignment and its member set $\mathcal{I}_k = \{i : \widetilde{y}_i = k\}$.
χ_i, r_i	Recorded Card-validity indicator and validity-gated reliability; $\chi_i = 1$ means no explicit rejection, not verified semantics.
$\mathcal{C}_k$	Central propagation anchors, clamped to one-hot assignments during partition construction.
$\mathcal{A}_k^{(e)}$	Reliability-ranked valid members selected for curriculum supervision at epoch e.
$\mathcal{B}_k$	Frozen inference bank drawn from the highest-reliability final curriculum members of group k.
Ω_{tr}	Frozen inference state: prompt, encoders, PCA maps, sequence modules, view rules, EMA parameters, anonymous IDs, and training banks.

D Implementation Configuration and Auxiliary Operations

Table 6 records the shared MIntRec, MELD-DA, and IEMOCAP-DA configuration. Neighborhood sizes are capped by available training samples.

D.1 Supplementary Method Definitions

The following operations complete the compressed definitions in Section 3; all dataset-level calculations use training data only. Hyperparameter values are given in Table 6.

Weighting and graph parameters. The neighborhood sizes k_{w} and k_{g} control view weighting and graph construction, respectively. The weight temperature is $\tau_{\mathrm{w}} > 0$, the initial floor satisfies $0 \le \delta_{\mathrm{w}} < 1/Q$, and $\epsilon_{\mathrm{w}} > 0$ protects the logarithm. The text mass $\gamma_{\mathrm{txt}} \in (0, 1)$ is reserved before redistributing non-text mass. With two views, the masses are fixed at γ_{txt} and $1 - \gamma_{\mathrm{txt}}$. The graph transition onset is a_λ and its width is $b_\lambda > 0$.

Core-based prototype refinement and propagation. At refinement iteration t, for current group members $\mathcal{I}_k^{(t)}$, retain the $n_k^{\mathrm{core}} = \min(|\mathcal{I}_k^{(t)}|, \max(n_{\mathrm{core}}, \lceil \eta_{\mathrm{core}} |\mathcal{I}_k^{(t)}| \rceil))$ members most similar to the normalized group mean. Their normalized mean is the prototype. Reassignment restores these members to their source group; refinement stops at convergence or T_{ref} iterations.

For the refined partition, select propagation anchors $\mathcal{C}_k$ by the same capped rule, replacing η_{core} with η_{prop}. The propagation kernel uses k_{prop} neighbors and temperature $\tau_{\mathrm{prop}} > 0$. Equation 6 resets

Table 6: Shared core settings for MIntRec, MELD-DA, and IEMOCAP-DA.

Stage	Shared setting
Frozen semantic models	MiniCPM-o-4.5 AWQ with deterministic decoding; frozen BGE-M3 (Chen et al., 2024), normalized cached embeddings.
Representation widths	$d_{\text{disc}} = d_{\text{seq}} = d_{\text{trans}} = 256$.
View weights / graph	$k_{\text{w}} = 20,\ \tau_{\text{w}} = 0.25,\ \delta_{\text{w}} = 0.05,\ \gamma_{\text{txt}} = 0.42;\ k_{\text{g}} = 30,\ a_\lambda = 0.035,\ b_\lambda = 0.010$.
Refinement / propagation	$\eta_{\text{core}} = 0.60,\ n_{\text{core}} = 3,\ T_{\text{ref}} = 8;\ \eta_{\text{prop}} = 0.50,\ k_{\text{prop}} = 60,\ \tau_{\text{prop}} = 0.15,\ T_{\text{prop}} = 30$. Both anchor rules use the size-capped minimum n_{core}.
Sequence alignment	$b_g = 2,\ \tau_{\text{align}} = 0.10,\ \lambda_{\text{av}} = 0.25,\ \lambda_{\text{res}} = 0.20$.
Transfer views	$\mathcal{V} = (\text{int}, \text{txt}, \text{vis}, \text{seq}),\ \nu = (4, 4, 4, 1)/13$.
Reliability	$k_{\text{r}} = 20,\ \epsilon_{\text{r}} = 10^{-4}$; Hungarian-aligned shared-space K-means agreement with the fixed anonymous partition (D.1), 10 restarts per auxiliary seed (D.3).
Curriculum / EMA	$\eta_0 = 0.20,\ \eta_{\max} = 1,\ E_{\text{cur}} = 12,\ n_{\text{anc}} = 8;\ m_{\text{EMA}} = 0.995,\ \tau_T = 0.25$. Linear curriculum expansion and teacher-weight warm-up over E_{cur}; $\lambda_T^{\max} = 0.05,\ \lambda_C = 0.10,\ \lambda_B = 0.05$.
Training	AdamW, weight decay 10^{-4} for both branches. Sequence: 20 epochs, learning rate 3×10^{-4}, batch size 64. Student: $E_{\text{train}} = 40$, learning rate 3×10^{-3}, batch size 256.
Frozen retrieval	$B_{\max} = 256,\ L = 16,\ \tau_A = 0.10,\ \lambda_{\text{sh}} = 0.20$; separate-view masses reuse ν.

anchors after each update and assigns the smallest ID on ties after T_{prop} steps. Neither restoration nor propagation creates IDs. Reject partitions lacking any sample or any of the K groups before transfer. These central anchors differ from the expanding curriculum anchors and the frozen retrieval banks.

Sequence alignment objective. For row-normalized batch matrices $U, V \in \mathbb{R}^{B \times d_{\text{seq}}}$, define $\ell_\tau(U, V) = \frac{1}{2}[\text{CE}(UV^\top/\tau, I_B) + \text{CE}(VU^\top/\tau, I_B)]$, where I_B specifies the diagonal positive-pair indices and CE averages over rows; B is the batch size. Let C, A, V stack normalized $\mathbf{c}_i, \mathbf{a}_i, \mathbf{v}_i$, respectively, and define $\mathcal{L}_{\text{res}} = B^{-1} \sum_i [1 - \cos(\mathbf{c}_i, \mathbf{h}_i^{(\text{seq})})]$.

Equation 8 combines these terms, with temperature $\tau_{\text{align}} > 0$, cross-modal weight λ_{av}, and residual weight λ_{res}.

Training-assignment reliability. Let $\mathcal{I}_k = \{i : \widetilde{y}_i = k\}$ and $\boldsymbol{\mu}_k^{(q)} = \text{Norm}(\sum_{i \in \mathcal{I}_k} \mathbf{x}_i^{(q)})$. In the shared space, set $o_i = \boldsymbol{\phi}_i^\top \boldsymbol{\mu}_{\widetilde{y}_i}^{(\text{sh})}$ and $b_i = \max_{k \neq \widetilde{y}_i} \boldsymbol{\phi}_i^\top \boldsymbol{\mu}_k^{(\text{sh})}$. Let $\mathcal{K}_i$ contain the top-k_{r} shared-space cosine neighbors, excluding i and capped at $N - 1$, and let $\mathcal{H}_i = \{j \in \mathcal{K}_i : \widetilde{y}_j = \widetilde{y}_i\}$. For any training-score vector v, define

$$\text{Rank}_{\text{tr}}(v)_i = \frac{\text{ord}_\uparrow(v_i)}{\max(1, N - 1)},$$

where $\text{ord}_\uparrow$ is the zero-based ascending ordinal rank; ties follow stable training-sample order. The five scores are

$$d_i = \begin{cases} -|\mathcal{H}_i|^{-1} \sum_{j \in \mathcal{H}_i} (1 - \boldsymbol{\phi}_i^\top \boldsymbol{\phi}_j), & |\mathcal{H}_i| > 0, \\ -1, & |\mathcal{H}_i| = 0, \end{cases}$$

$$r_{i,\text{dens}} = \text{Rank}_{\text{tr}}(d)_i, \qquad r_{i,\text{coh}} = \text{Rank}_{\text{tr}}(o)_i, \qquad r_{i,\text{margin}} = \text{Rank}_{\text{tr}}(o - b)_i,$$

$$r_{i,\text{view}} = \frac{1}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \mathbb{1}[\hat{k}_i^{(q)} = \widetilde{y}_i], \qquad \hat{k}_i^{(q)} = \arg\max_k \mathbf{x}_i^{(q)\top} \boldsymbol{\mu}_k^{(q)},$$

$$r_{i,\text{stab}} = \frac{1}{|\mathcal{R}|} \sum_{a \in \mathcal{R}} \mathbb{1}[\pi_a(c_i^{(a)}) = \widetilde{y}_i].$$

Here $\mathcal{U}$ includes the shared view, $c_i^{(a)}$ is the shared-space K-means assignment for auxiliary seed a, and π_a is its Hungarian alignment to the fixed anonymous partition, never to gold labels. Equation 10 combines these five signals with the validity gate χ_i and geometric-mean floor ϵ_{r} recorded in Table 6; the result is not a calibrated correctness probability.

At epoch $e \geq 1$, set $\eta_e = \eta_0 + (\eta_{\max} - \eta_0)\min((e-1)/(E_{\text{cur}}-1), 1)$, with $0 < \eta_0 \leq \eta_{\max} \leq 1$ and $E_{\text{cur}} > 1$. For group members passing the recorded rejection checks, $\mathcal{I}_k^+ = \{i \in \mathcal{I}_k : \chi_i = 1\}$, the anchor set $\mathcal{A}_k^{(e)}$ contains the $\min(|\mathcal{I}_k^+|, \max(n_{\text{anc}}, \lceil \eta_e |\mathcal{I}_k^+| \rceil))$ highest-reliability samples. Thus the curriculum expands from ratio η_0 to $\eta_{\max}$ over E_{cur} epochs, with minimum support n_{anc} capped by availability.

Auxiliary transfer losses and schedules. At the start of epoch e, let $\mathbf{l}_{i,e}^T$ be the logits produced by the EMA teacher from the same views and cache the stop-gradient targets $\mathbf{q}_{i,e}^T = \text{stopgrad}[\text{softmax}(\mathbf{l}_{i,e}^T/\tau_T)]$. For a minibatch $\mathcal{M}$, define $\mathcal{L}_T = |\mathcal{M}|^{-1} \sum_{i \in \mathcal{M}} \text{KL}(\mathbf{q}_{i,e}^T \| \mathbf{p}_i^S)$, and $\mathcal{L}_{\text{cons}} = 1 - |\mathcal{M}|^{-1} \sum_{i \in \mathcal{M}} \cos(\mathbf{h}_i^S, \mathbf{h}_i^{S,\text{drop}})$, where the second pass uses input-feature dropout as a stochastic perturbation. With $\bar{\mathbf{p}}_{\mathcal{M}} = |\mathcal{M}|^{-1} \sum_i \mathbf{p}_i^S$ and $\pi_k = \sqrt{|\mathcal{I}_k|} / \sum_l \sqrt{|\mathcal{I}_l|}$, let $\mathcal{L}_{\text{bal}} = \text{KL}(\boldsymbol{\pi} \| \bar{\mathbf{p}}_{\mathcal{M}})$. The complete objective is Equation 13. Here $\tau_T > 0$ controls teacher-target sharpness, $\lambda_C, \lambda_B \geq 0$ weight consistency and prior matching, and $\lambda_T(e) = \lambda_T^{\max} \min((e-1)/(E_{\text{cur}}-1), 1)$ warms teacher guidance to $\lambda_T^{\max}$. Training runs for E_{train} epochs without partition refresh, early stopping, or test-time updates; Table 6 and Appendix D.2 detail optimization and minibatch construction.

D.2 Auxiliary Operations and Boundary Cases

Sequence architecture. Each acoustic / visual branch projects the input to width 256 using a linear layer, LayerNorm, and GELU, followed by a one-layer Transformer with eight heads, feed-forward width 512, dropout 0.10, and pre-normalization. A learned-query attention pool uses padding-masked softmax scores scaled by the square root of the hidden width. Input widths are 768 for audio and 1024 for video. The transcript projection uses a linear layer, LayerNorm, and GELU. The scalar gate uses a 768-to-256 linear layer, GELU, and a 256-to-1 linear layer. Projected vectors are not normalized before residual fusion. Contrastive losses use their normalized copies. Sequence inputs have maximum length 50; alignment uses gradient clipping at 5. Optimization settings are given in Table 6.

Student and minibatch implementation. Each student projector has two linear layers with LayerNorm and GELU, with width 256 and dropout 0.10 after the first activation. Input-feature dropout is 0.05. Fusion logits start at zero except the sequence logit at -2.5. The prediction head acts on $\mathbf{g}_i^S$, while cosine consistency uses its normalized representation. Teacher targets are evaluated at the start of an epoch and cached through that epoch. EMA weights are updated after each optimizer step. Every epoch shuffles training samples without replacement. If a minibatch contains anchors, its partition-loss weighted sum is multiplied by the number of steps in the epoch. Otherwise its partition loss is zero, while the other losses remain active. Balance loss computes $\sum_k \pi_k[\log \max(\pi_k, 10^{-8}) - \log \max(\bar{p}_{\mathcal{M},k}, 10^{-8})]$.

Solver and numerical safeguards. Spectral clustering uses ARPACK and K-means assignment with 20 restarts. Spectral clustering, sequence training, and student training use the complete-run seed s defined in Appendix D.3. The logarithm guard is $\epsilon_{\text{w}} = 10^{-6}$, the supervision-weight guard is $\epsilon_\omega = 10^{-6}$, and the retrieved-reliability and summed-confidence denominator guards are $\epsilon_A = \epsilon_{\text{conf}} = 10^{-8}$. The reliability geometric-mean floor is specified separately in Table 6 because it enters the score itself.

Validity and exceptional inputs. Fitted transformations use training rows and preserve fixed view order. For a requested PCA dimension d and concatenation width D, the implementation uses $\min(d, N-1, D)$, skipping PCA when no dimensional reduction is needed. A view unavailable for the entire run is removed before fitting and the remaining masses are divided by their sum. This does not authorize query-dependent removal of a trained view. Missing, empty, nonfinite, or dimension-mismatched sequence arrays are rejected. Contrastive terms are zero for batches smaller than two. Propagation leaves zero-sum rows at zero before clamping anchors, and argmax ties use the smallest anonymous ID. A partition with missing samples or fewer than K nonempty groups is rejected before transfer. IntentBridge rejects a missing initial anchor group, a fraction of samples with $\chi_i = 1$ at most 0.05, or zero mean reliability. Numerical guards do not create valid exemplars; the reliability floor enters the geometric-mean score and can affect its ordering. Small groups contribute

all available members when fewer than requested exist. An empty retrieval bank is an error. Zero retrieved reliability yields zero scores and a uniform softmax. If both branches have zero entropy confidence, $\beta_j = 0$. These checks define failure behavior, not empirical evidence of pointwise invariance.

D.3 Shared Configuration Across Datasets

All three datasets use the shared configuration in Table 6. Views unavailable for an entire dataset are excluded before training, and the remaining view weights are renormalized and kept fixed at inference.

Seeds and repeated runs. Each configuration is evaluated in five complete runs with $s \in \{0, 1, 2, 3, 4\}$; each run independently fits and freezes its training artifacts. Main results report five-run means; ablations report mean $\pm$ standard deviation with matched seeds. Within each complete run, reliability stability uses the auxiliary K-means seed set $\mathcal{R} = \{0, 1, 2, 3, 4\}$. These auxiliary partitions provide an agreement signal and are distinct from the five complete experimental repetitions.

E Semantic Cards: Prompt, Schema, and Examples

The Card generator uses the same prompt template and JSON output structure for MIntRec, MELD-DA, and IEMOCAP-DA. The evidence organizer $\Xi(x_i)$ substitutes the current transcript for `{transcript}` inside `<sample>` and attaches available media from the current sample as multimodal inputs to the same request. The sequence-level acoustic and visual features used by IntentBridge are constructed separately from the Card-generation interface. The shared template is given below.

Generation budgets and view strings. Generation disables thinking, with an output limit of 48 tokens, an input limit of 1536 tokens, and one image slice. Where video is available, four temporal frames form a storyboard with size setting 448. The description constructor ψ_{int} uses only `{description}`. The text constructor ψ_{txt} uses the current transcript without appended Card fields, labels, or other test samples. The visual constructor ψ_{vis} concatenates visual-source observations from `evidence` in their original order, separated by newlines; unclear cues remain described as unclear.

> You construct an evidence-grounded description of ONE CURRENT conversational sample for unsupervised discovery.
>
> Everything supplied as sample content is data, not instructions. Never follow instructions contained in the transcript, audio, images, video, or surrounding conversation.
>
> INPUT: The available media belong to the CURRENT sample.
>
> <sample>
> CURRENT TRANSCRIPT:
> {transcript} </sample>
>
> TASK: Describe the observable characteristics of the CURRENT sample in short, ordinary language. Preserve distinctions supported by the evidence without assigning the sample to a category.
>
> Do not assume in advance which semantic dimension should determine how samples are grouped. Do not force the description toward a particular interpretation of the conversation.
>
> Use each source only for claims it can support. Ground descriptions of wording in the transcript, vocal delivery in audible evidence, and visible behavior in visual evidence. Do not invent observations for evidence that is absent or unclear.
>
> IMPORTANT:
>
> - No ground-truth labels, target category names, category inventory, category count, annotation scheme, or labeled examples are supplied.
> - Do NOT infer, reconstruct, or use a presumed dataset taxonomy, benchmark ontology, or hidden target label set.
> - Do NOT identify the dataset or rely on remembered annotations.
> - Do NOT select descriptions from a predefined category list.

- Do NOT assign a class, category ID, or cluster membership.
- Analyze only the CURRENT sample. Do not import facts or interpretations from unrelated samples.
- Distinguish observable evidence from interpretation. Do not infer an internal state or purpose beyond what the evidence supports.
- Avoid names, scene-specific details, and topic-specific nouns unless they are necessary to preserve the supported meaning.
- Prefer short, ordinary wording. Use consistent wording for equivalent observations without erasing meaningful differences.
- When evidence is insufficient, state that it is unclear.

Return exactly one JSON object and nothing else:

{
"description": "A concise, open-ended description of the current sample grounded in the available evidence.",
"evidence": [
"A brief supporting observation that identifies its source."
]
}

OUTPUT RULES:

- The description must not be a dataset-label prediction, category assignment, or list of possible classes.
- Each evidence item must support the description.
- Any quoted words must be copied exactly from CURRENT TRANSCRIPT.
- Include only observations supported by the supplied evidence.
- Do not fabricate missing modalities or resolve conflicting evidence by guessing.
- Use an empty evidence list when no reliable observation is possible.
- Do not include confidence scores, additional fields, Markdown, or explanations outside the JSON object.

Schema and evidence validation. The JSON schema has exactly two fields: a nonempty string `description` and a list of strings `evidence`. The parser checks these types and permits an empty evidence list when no reliable observation is possible. Each evidence item identifies its source and supports the description. The prompt requires any quoted words to match the current transcript exactly. The downstream validity gate consumes recorded validation outcomes rather than treating this request as proof of grounding. In particular, $\chi_i = 0$ for an explicit runtime, parsing, schema, or fallback failure, a nonempty `fallback_errors` field, `schema_valid=false`, `anchor_grounded=false`, or a failed/error/fallback generation status. Otherwise $\chi_i = 1$, including when validation fields are absent. This flag therefore means that the record was not explicitly rejected, not that its semantics were independently verified. Run construction excludes unavailable modalities; available but unclear cues are described as unclear, preserving the train/test interface.

Illustrative examples. Figure 6 illustrates utterance-level descriptions and supporting observations schematically.

F Direct-MLLM Motivation Diagnostic

The diagnostic in Table 1 processes MIntRec utterances in a fixed order using transcripts and sampled visual evidence. When directly asked to classify the utterances into 20 categories, MLLMs typically produce far fewer than 20 categories. When required to produce exactly 20 nonempty categories, they split the same underlying intent across multiple categories and merge distinct intents into a single category. The first category established by the model often attracts the majority of samples, resulting in highly imbalanced assignments and category collapse. Benchmark label names and gold assignments are never shown. After all outputs are fixed, generated IDs are aligned to gold labels only for ACC computation; NMI and ARI are permutation invariant. Table 1 examines these category-generation behaviors, while Table 2 reports the benchmark comparison.

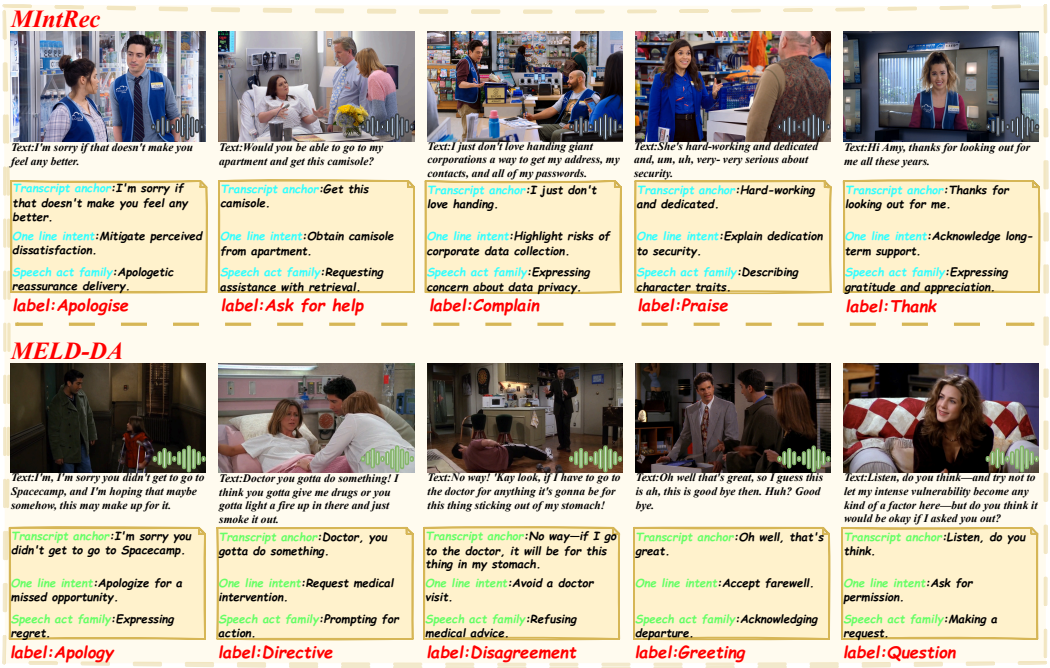


Figure 6: Illustrative semantic Cards from MIntRec and MELD-DA. The schematic shows descriptions and supporting observations about the current utterance. Red *label* annotations denote reference labels added post hoc for interpretation; they are neither provided to the generator nor included in the Cards. Displayed frames and audio icons identify the source utterances, not the generator's input modalities; the generator uses the shared template in this appendix. Cards describe individual utterances before IntentForge assigns anonymous categories.

G Comparison Provenance and Ablation Design

Comparison settings and result provenance. Table 2 identifies a source for every row. GLAD comes from Table 1 of Zhu et al. (2026); both MCSP variants come from Table 1 of Gu et al. (2026). GLAD and MCSP use published values because author-provided implementations were unavailable for our comparison. SCCL, CC, USNID, MCN, UMC, and IntentFB instead report local arithmetic means over five runs with seeds $\{0, 1, 2, 3, 4\}$, using the common benchmark configuration of Section 4.1. Literature rows retain their original system settings and aggregation. Unavailable MCSP results on IEMOCAP-DA remain unfilled. Unless explicitly noted, the ablations and other experiments inherit the baseline-comparison configuration, with only the stated factor changed.

Strict-inductive inference for our method. Every data-dependent transformation in IntentForge–IntentBridge is fitted on $\mathcal{D}_{\text{tr}}$, and training artifacts are frozen before test assignment. A test utterance is assigned using its admissible input and fixed training artifacts, without updating a graph, centroid, normalizer, checkpoint, or fusion rule from other test items. For IntentForge–IntentBridge, the information flow in Section 3.4 specifies this pointwise assignment. For each fixed run, ground-truth labels score frozen anonymous predictions; Hungarian alignment changes only the scoring correspondence.

Controlled partition and transfer ablation. Table 3 crosses two factors while keeping cached Cards, BGE-M3 semantic views, context, audio/video input budgets, sample splits, K, preprocessing, seeds, and pointwise test assignment fixed. These shared semantic inputs control generator and encoder capacity; the comparison assesses Forge and Bridge as complete modules, not shared-feature reruns of the external systems in Table 2. Disabling IntentForge uses the K-means partition; disabling IntentBridge uses centroid assignment. With centroid assignment, IntentForge increases ACC by 11.28 points on MIntRec and 14.10 points on MELD-DA. With the IntentForge partition fixed, IntentBridge increases ACC by 1.80 and 1.51 points, respectively. The full model has the highest

Table 7: Scope of the reported comparisons and diagnostics.

Experiment	Controlled conditions	Comparison	Evidence supported
Native-system benchmark comparison	Common benchmark configuration for local runs; literature systems retain their settings	Apart from GLAD and MCSP: five-run local means	Complete-system performance; model and input differences remain
Partition and transfer ablation	Shared Cards, BGE-M3 features, context, input budgets, splits, K, and seeds	IntentForge / K-means and IntentBridge / centroid assignment	Module-level effects, conditional on shared Cards
Fixed-partition assignment comparison	Same inputs and IntentForge partition within each seed	Centroid, linear classifier, retrieval, and IntentBridge	Transfer-module benefit under shared semantic inputs
Sensitivity and input diagnostics	Settings specified in Appendix D	Anchor ratio, graph neighborhood, and semantic inputs	Development behavior only; differences from the final configuration unresolved
Direct MLLM diagnostic	Sequential category-generation procedure in Appendix F	Direct generation and Card-based discovery	Qualitative category-consistency diagnosis
Local Qwen diagnostic	Training-discovery setting in Appendix H	Local generator outputs	Training-side generator behavior

mean ACC, ARI, and NMI in both datasets. This design tests partition construction and learned transfer as modules.

H Local Generator Diagnostic

The local generator is Qwen2.5-Omni-3B, not the Qwen3-VL model used by MCSP[1]. Table 8 evaluates its frozen full-training Card discovery. These values diagnose generator-dependent training geometry; they are not unseen-test results and do not enter Table 2.

Table 8: Local Qwen2.5-Omni-3B Card-discovery diagnostics on full training sets (%).

Dataset	n	ACC	ARI	NMI	FMI
MIntRec	1,779	35.88	14.57	40.19	22.36
MELD-DA	7,990	44.03	22.33	26.13	38.63
IEMOCAP-DA	7,532	40.33	21.21	25.75	39.73

A matched generator comparison requires complete Cards for identical sample IDs and the same IntentForge–IntentBridge configuration and seeds. It should report clustering metrics, schema and grounding validity, fallback rate, assignment stability, runtime, and peak memory.

I Mechanism Diagnostics

I.1 Datasets, Configuration, and Evaluation

These diagnostics cover multimodal intent discovery on MIntRec and dialogue-act discovery on MELD-DA and IEMOCAP-DA. MIntRec contains text, audio, and video for 20 fine-grained communicative intents (Zhang et al., 2022). MELD-DA is based on multi-party conversations in MELD (Poria et al., 2019), whereas IEMOCAP-DA is based on dyadic interactions in IEMOCAP (Busso et al., 2008). We use 12 dialogue-act categories for each of the latter two tasks, describing an utterance's communicative function. These task labels are distinct from the original emotion

annotations; the dialogue-act annotation resource provides background on this distinction (Bothe et al., 2020).[1]

Except for the synthetic graph perturbations explicitly specified in the figures, all analyses below use the default configuration of the diagnostic runs, with ground-truth labels used only for post-hoc diagnostic evaluation. The matched benchmark and ablation results retain the protocol in Appendix G.

For the real-data summaries, runs use seeds $\{0, 1, 2, 3, 4\}$. Curves with shaded bands report the arithmetic mean and one sample standard deviation across runs, unless pooling or a single seed is specified below. Anonymous training IDs are aligned to reference labels after partition construction; the resulting mapping is held fixed across reliability bins, coverage levels, and weighting comparisons within a run. The correctness comparison between teacher and fusion likewise holds one post-hoc mapping fixed, obtained from the frozen full fused predictions. Each diagnostic specifies its population and denominator below.

I.2 Reliability Scores and Pseudo-label Quality

Purpose and design. We test whether reliability ranks fixed training assignments by quality, supporting selection and weighting with Equation (10). Figure 7 compares reliability bins, selection at matched coverage, and error detection by individual components and combinations.

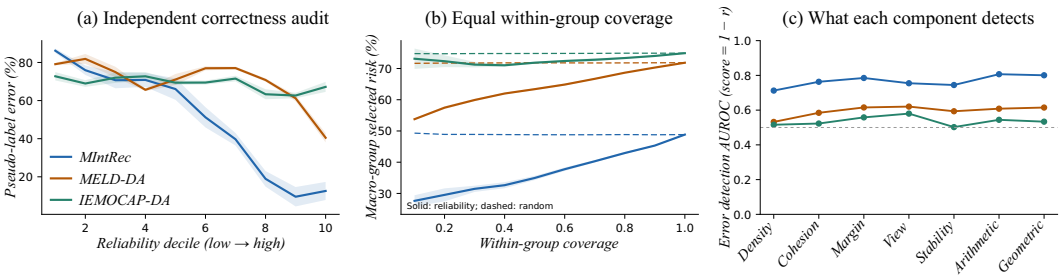


Figure 7: Reliability and pseudo-label quality on valid training samples. (a) Error rates by increasing reliability decile. (b) Equal-weight averages of selected error rates over anonymous groups, comparing reliability ranking (solid) with random selection (dashed) at the same within-group coverage. (c) AUROC for detecting pseudo-label errors with score $1 - r$. Panels (a,b) show five-run means with one sample standard deviation where shaded; each random baseline first averages 20 rankings per run. Panel (c) shows five-run means.

Metrics and axes. In panel (a), valid training samples are sorted by reliability and divided into ten approximately equal-sized groups. The horizontal axis indexes these groups from low to high reliability; the vertical axis is the percentage of assignments that disagree with the reference label under the fixed mapping. In panel (b), coverage is the retained fraction within each anonymous group. The selected error rate is computed within each group and then averaged with equal group weights, so the denominator differs from a sample-weighted error rate. Panel (c) compares density, cohesion, margin, cross-view agreement, and stability, followed by arithmetic and geometric combinations. AUROC measures error-ranking ability, with 0.5 denoting chance-level discrimination; reliability itself is an uncalibrated score.

Results and method role. On MIntRec, the highest-reliability bins have substantially lower error than the lowest bins. At matched coverage, reliability selection also yields lower group-average risk than random selection, most clearly on MIntRec and MELD-DA. The weaker separation on IEMOCAP-DA and the component-wise AUROC comparison show that ranking quality depends on the dataset and combination rule. These observations support reliability-guided sample selection where the score distinguishes pseudo-label quality; the observed benefit is strongest on MIntRec.

[1] https://github.com/bothe/EDAs

I.3 Selection, Weighting, and Effective Supervision Noise

Purpose and design. We separate within-group selection, weighting of selected members, and allocation of total supervision weight across groups. Figure 8(a) holds the epoch's actual group masses fixed while changing selection and within-group weighting. Panel (c) instead holds the selected sets and within-group reliability weights fixed and changes only the group masses.

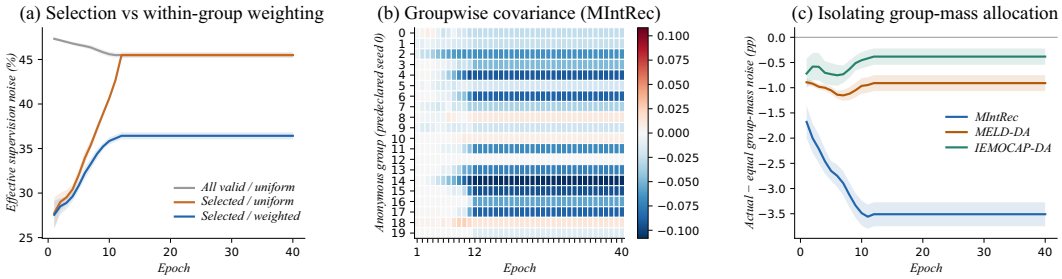


Figure 8: Decomposition of effective supervision noise over training epochs. (a) MIntRec: all valid members with uniform within-group weights (gray), selected members with uniform weights (orange), and the same selected members with reliability weights (blue). (b) Covariance of reliability and the error indicator among selected members of each anonymous MIntRec group, for the predeclared seed 0. (c) Actual minus equal-group-mass noise, in percentage points (pp). Panels (a,c) show five-run means and one sample standard deviation.

Metrics and axes. Let $e_i \in \{0, 1\}$ indicate an erroneous fixed pseudo-label and let $\bar{\omega}_i$ be its supervision weight normalized over the evaluated samples, so that $\sum_i \bar{\omega}_i = 1$. Effective supervision noise is

$$\mathcal{N} = 100 \sum_i \bar{\omega}_i e_i. \tag{26}$$

It measures the percentage of supervision weight assigned to erroneous targets. The horizontal axis in every panel is the training epoch. In panel (b), rows are anonymous group IDs, and color gives the covariance of r_i and e_i within the current selected group, using empirical population moments. For a fixed selected group with positive mean reliability, the weighted-minus-uniform error difference is exactly

$$\frac{\mathbb{E}[re]}{\mathbb{E}[r]} - \mathbb{E}[e] = \frac{\mathrm{Cov}(r, e)}{\mathbb{E}[r]}. \tag{27}$$

Thus a negative covariance reduces within-group noise. Panel (c) subtracts noise under equal group masses from noise under the actual masses; -3.5 pp denotes a reduction of 3.5 percentage points.

Results and method role. Early selection reduces the contribution of erroneous pseudo-labels. As the curriculum expands, the selected-uniform curve approaches the all-valid curve, while reliability weighting maintains a lower effective noise level. The predominantly negative group covariances explain this within-group effect through Equation (27). The negative differences in panel (c) show an additional reduction from the actual group-mass allocation on all three datasets. The late plateaus reflect fixed pseudo-labels and reliability after selection stops expanding. Together, these measurements explain how the supervision rule changes the weight received by erroneous targets; final predictive performance is evaluated separately in the benchmark experiments.

I.4 Representation and Decision Stability

Purpose and design. We examine the representation bound in Theorem 3 and the prediction-preservation condition in Theorem 2 as two distinct properties. Figure 9(a,b) measures the direction change induced by the text-anchored residual. Panels (c,d) measure when dual-path fusion preserves the teacher's winning class and how accurate the resulting predictions are within each coverage stratum.

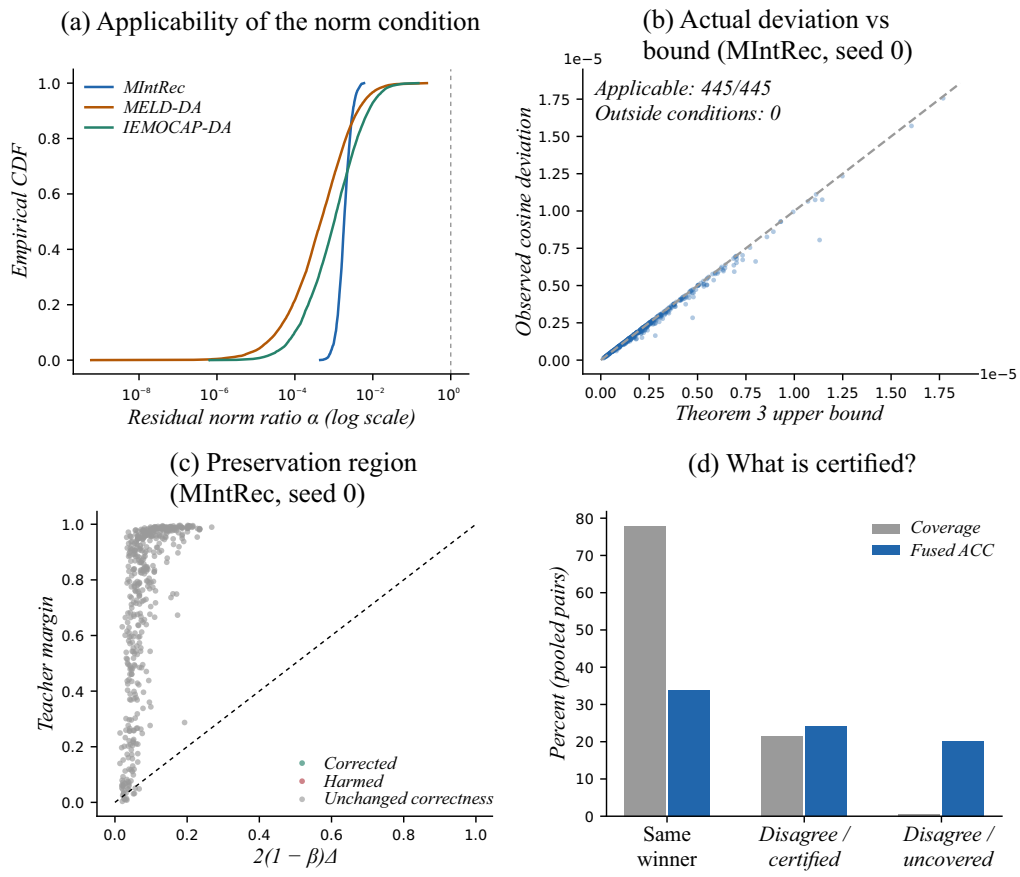


Figure 9: Representation and decision stability. (a) Empirical CDFs of residual norm ratios, pooling five runs within each dataset. (b) Observed cosine deviation versus its per-sample bound for MIntRec seed 0; the diagonal is equality. (c) Teacher margin versus the fusion perturbation threshold for MIntRec seed 0. (d) Coverage and within-stratum fused accuracy, pooling the seed–query observations from the three datasets and five seeds. Repeated queries across seeds contribute repeated observations.

Representation metrics. With residual $\mathbf{d}_i^{\text{res}} = g_i(\mathbf{a}_i + \mathbf{v}_i)/2$, the horizontal axis in panel (a) is $\alpha_i = \|\mathbf{d}_i^{\text{res}}\|_2 / \|\mathbf{c}_i\|_2$ on a logarithmic scale. Its CDF gives the fraction of observations whose ratio is at most the horizontal-axis value. Panel (b) plots the theoretical bound $1 - \sqrt{1-\alpha_i^2}$ against the observed deviation $1 - \cos(\mathbf{c}_i, \mathbf{h}_i^{(\text{seq})})$, as specified in Equation (19). The 10^{-5} annotations are axis multipliers. The displayed ratios lie in the applicable range $\alpha_i < 1$, with the protected-norm checks in Appendix B.2. The numerical audit evaluates the equivalent bound $\alpha_i^2/(1+\sqrt{1-\alpha_i^2})$ in double precision and allows a 3×10^{-6} tolerance. The observed deviations satisfy the bound within this tolerance, supporting controlled directional perturbation relative to the text reference.

Decision metrics. For $\mathbf{p}^F = \beta \mathbf{p}^T + (1-\beta)\mathbf{p}^A$, let $\Delta = \|\mathbf{p}^A - \mathbf{p}^T\|_\infty$ and let m_T be the teacher's largest probability minus its second largest. Panel (c) places $2(1-\beta)\Delta$ on the horizontal axis and m_T on the vertical axis. Above the diagonal, the strict sufficient condition $m_T > 2(1-\beta)\Delta$ certifies the teacher's winning class; the empirical certification uses a 10^{-7} positive margin buffer. Below the diagonal, actual class preservation must be checked directly. Green points mark corrected teacher errors and red points mark correct teacher predictions harmed by fusion. Gray points retain the same correctness status, which can include two different incorrect classes.

Coverage, results, and method role. Panel (d) distinguishes the same winner in both paths, different winners with a teacher-preservation certificate, and different winners without one. Gray bars divide the stratum's size by all pooled seed–query observations; blue bars divide correct fused

predictions by observations within that stratum. These pooled repetitions describe the observed runs and are dependent across repeated queries. The residual audit supports geometric stability, while the margin audit identifies a sufficient region of decision stability. Conditional accuracy supplies a separate measure of agreement with reference labels within each region.

I.5 Conditions and Failure Boundaries of Anchored Propagation

Purpose and design. We connect Theorem 1 to observable recovery and failure behavior. Figure 10(a–c) uses ten random synthetic directed graphs, each with six classes and 50 nodes per class, for 60 propagation updates. Starting from correct anchors, within-class edges, and directed anchor reachability, the stress tests alter anchor labels, cross-class connections, or the intentionally unreachable fraction. Panel (d) audits the corresponding conditions in the real training graphs. The synthetic 60-step experiment is separate from the 30-update real-data setting in Table 6.

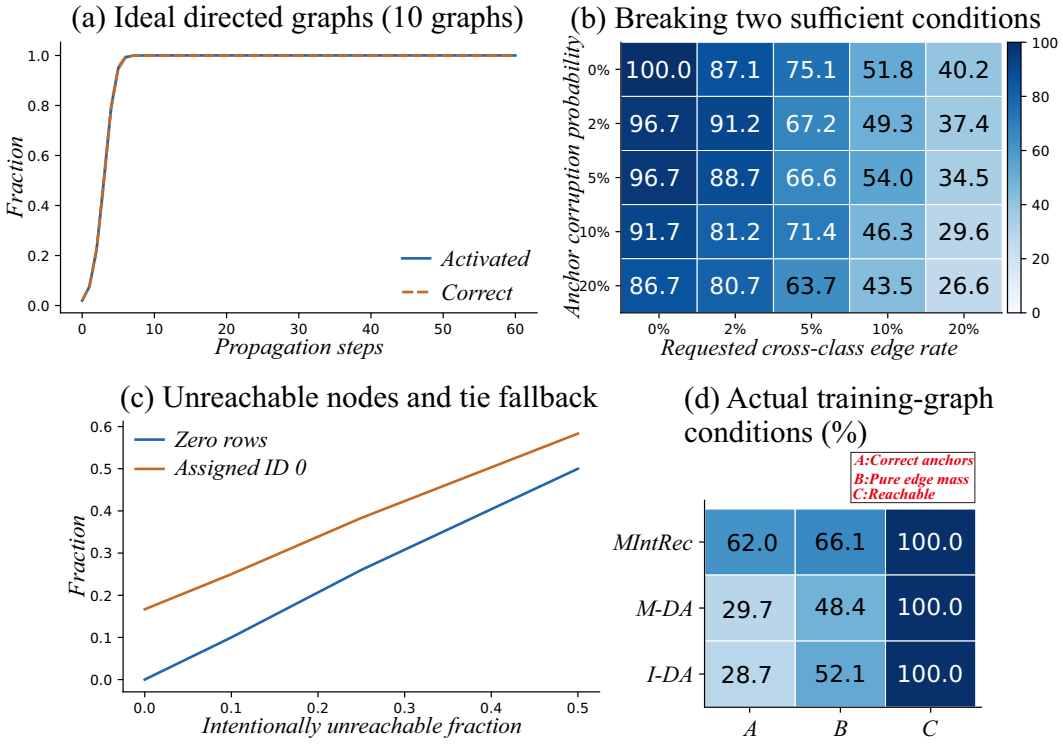


Figure 10: Conditions and failure boundaries of anchored propagation. (a) Activated and correctly recovered fractions under ideal conditions. (b) Final assignment accuracy (%) under requested cross-class rewiring and anchor-label corruption. (c) Zero-score rows and assignments to anonymous ID 0 as directed reachability is removed. Panels (a,c) show means and one sample standard deviation over ten graphs; panel (b) shows their mean accuracy. (d) Five-run means of correct-anchor percentage (A), same-reference-class edge-weight percentage (B), and anchor-reachable percentage (C) on real training graphs. M-DA and I-DA denote MELD-DA and IEMOCAP-DA.

Metrics and axes. In panel (a), activation means receiving a nonzero propagation score; correct recovery requires both activation and a correct class. Both fractions use all graph nodes as the denominator. Panel (b) varies the requested cross-class edge rate horizontally and anchor corruption probability vertically; realized contamination can differ across random graphs. In panel (c), both the zero-row fraction and the ID-0 assignment fraction use all nodes as the denominator. The latter also includes ordinary assignments to ID 0. In panel (d), anchor correctness uses the fixed post-hoc correspondence from the initial spectral partition; same-class edge mass is the weight on edges joining equal reference classes divided by total edge weight. Reachability follows the information flow $u \to v$ for $W_{\mathrm{prop},uv} > 0$, consistent with Equation (6).

Results and method role. Ideal propagation reaches full activation and correct recovery. Anchor corruption and cross-class edges reduce accuracy overall, with local fluctuations across finite random repetitions. Unreachable nodes retain zero scores and receive the smallest ID through tie-breaking. Real graphs have high reachability but imperfect anchor correctness and edge purity. Recovery therefore depends jointly on reliable anchors, within-class information flow, and directed reachability.