SEMI-SUPERVISED MULTIMODAL NEW INTENT DISCOVERY VIA PROGRESSIVE LEARNING AND MLLM-GUIDED SEMANTIC CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Semi-supervised new intent discovery aims to identify unseen intents using labeled examples of known intents and unlabeled data. Existing methods primarily rely on text and overlook complementary acoustic and visual information. Moreover, semantic feedback from large language models can preserve incorrect intent groupings when conveyed through summaries of the current clusters or local rewriting. We propose PLSC, a method that combines progressive multimodal representation learning with semantic correction guided by a multimodal large language model (MLLM). To mitigate instability in semi-supervised multimodal clustering when modality cues are inconsistent, the progressive strategy stabilizes shared multimodal representations through supervised contrastive learning using known intent labels before each clustering round and gradually increases the contribution of pseudo-label classification. To correct erroneous intent groupings, we incorporate local semantic relations inferred by the MLLM from representatives as soft constraints in centroid refinement, allowing the corrected centroids to guide subsequent assignments and optimize intent boundaries. Experiments on MIntRec, MIntRec2.0, and MELD-DA demonstrate consistent gains of approximately 1% to 3% over existing methods across different known class ratios, providing a methodological foundation for semi-supervised multimodal clustering.

1 INTRODUCTION

Semi-supervised multimodal new intent discovery (NID) seeks to identify previously unknown intents by leveraging a small set of examples from known intents and a large pool of unlabeled multimodal instances. By jointly exploiting textual, acoustic, and visual information, it classifies instances associated with known intents and groups those expressing new intents into coherent clusters, thereby expanding the intent taxonomy while reducing the need for manual annotation. This task is relevant to applications including multimodal conversational agents, human–robot interaction, customer-service systems, in-vehicle assistants, and multimodal understanding of instructional content and livestreams, where text alone often fails to fully capture user intent (Tiwari et al., 2022; Lukin et al., 2024; Okur et al., 2020; Maharana et al., 2022). In these applications, semi-supervised multimodal NID provides a practical approach for improving intent understanding in open-world environments, reducing annotation costs, and facilitating adaptation to newly emerging intents.

Research on semi-supervised NID has shifted from representation learning for clustering toward integrating semantics derived from large language models into the discovery process. Early studies formulated semi-supervised NID mainly as a clustering problem, using labeled examples of known intents to shape the representation space and refining assignments for unlabeled utterances through clustering, iterative assignment refinement, contrastive learning, and probabilistic estimation (Lin et al., 2020; Zhang et al., 2021b; 2022b; Zhou et al., 2023). More recently, research on NID across different supervision settings has leveraged large language models to generate intent descriptions, summarizing groups of utterances, or providing semantic feedback to the NID model (De Raedt et al., 2023; Rodriguez et al., 2024; Fan et al., 2024). However, existing methods suffer from two limitations. First, prior semi-supervised NID work has remained text-only, leaving acoustic and visual cues unused even when they are essential for distinguishing lexically similar utterances with various intents, as illustrated in Appendix B.1. Second, LLM-based methods such as NILC con-

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.

dition semantic feedback on the current partition and incorporate it through cluster summaries or local rewriting. This dependence may reinforce incorrect intent groupings, leaving distinct intents conflated or instances of the same intent split across clusters, as analyzed in Appendix B.2.

To address these limitations, we propose PLSC, a semi-supervised multimodal new intent discovery method that combines progressive multimodal learning with semantic centroid correction. To exploit complementary modality cues under limited supervision, the progressive strategy addresses the instability in representation learning and cluster assignments that can arise from inconsistencies across modalities. Building on multimodal pretraining, it reinforces the semantic structure of the shared space through a lightly weighted supervised contrastive update using known intent labels before each clustering round, while consistency across views promotes representation alignment. As the subset selected by density expands, the weight of pseudo label classification increases gradually up to a small upper bound, allowing discovered assignments to provide stronger supervision while reducing the risk of amplifying early assignment errors. To correct erroneous intent groupings, we introduce semantic centroid correction guided by a multimodal large language model (MLLM). We select the top M representatives from each candidate new cluster using the margin criterion of MCSP (Gu et al., 2026) and obtain semantic judgments from their text and video evidence about whether representative pairs express the same intent. These judgments serve as soft penalties in a pairwise constrained k-means (Basu et al., 2004) objective that jointly refines new intent centroids and representative assignments while keeping known intent centroids fixed at the mean representations of their labeled samples during correction. The corrected centroids initialize the next clustering round so that local semantic relations guide subsequent assignments and refine intent boundaries.

We summarize our contributions as follows. (1) To the best of our knowledge, PLSC is the first multimodal method for semi-supervised new intent discovery, with a progressive learning strategy that exploits complementary modality cues while reducing the effect of early assignment errors. (2) We extend the role of LLM-based semantics in semi-supervised NID by using semantic relations to correct cluster centroids, allowing semantic distinctions between representative samples to optimize intent boundaries. (3) Experiments on MIntRec, MIntRec2.0, and MELD-DA demonstrate consistent gains of approximately 1% to 3% over existing methods across different known class ratios, establishing PLSC as a principled method for semi-supervised multimodal clustering.

2 Related Work

2.1 Semi-supervised Intent Discovery

Semi-supervised NID leverages a small labeled set of known intents to guide the discovery of both known and new intent categories in unlabeled data. Early supervised methods injected this supervision directly into the clustering process. CDAC+ (Lin et al., 2020) combined pairwise constraints with iterative target-distribution refinement, whereas DeepAligned (Zhang et al., 2021b) used supervised pre-training and aligned cluster identities across iterations to stabilize assignments. Research subsequently shifted toward learning transferable representations that more reliably generalize beyond known classes. SCL (Shen et al., 2021) and MTP-CLNN (Zhang et al., 2022b) incorporated labeled supervision into contrastive objectives, whereas CoCoID (Cao et al., 2022) and neighborhood-based methods (Mou et al., 2022) exploited local consistency and cluster compactness to mitigate bias toward known intents. Subsequent efforts improved pseudo-label reliability by addressing assignment inconsistency, forgetting of prior knowledge, and boundary ambiguity. USNID (Zhang et al., 2024c) stabilized training through historical centroid initialization, while LatentEM (Zhou et al., 2023) used EM-based latent-variable inference to preserve prior knowledge. The approaches of Mou et al. (2023) and Shi et al. (2023) further refined pseudo-labels through mutually reinforcing prototype disambiguation and structure-aware graph smoothing, respectively. Additional related work on unsupervised intent discovery and multimodal language analysis is provided in Appendix C.1 and Appendix C.2 for extended coverage and further discussion.

2.2 New Intent Discovery with Large Language Models

Large language models have expanded NID from embedding-space organization to explicit linguistic reasoning. One line of work treats LLMs as the discoverer itself, as exemplified by early evaluations of ChatGPT for open-world discovery (Song et al., 2023) and by IntentGPT (Rodriguez et al.,

2024), which uses few-shot prompting to propose and update intent inventories without additional model training. A complementary direction converts LLM outputs into supervision for smaller models, where IDAS (De Raedt et al., 2023) abstracts utterances prior to clustering, LANID (Fan et al., 2024) extracts pairwise semantic relations, and SynCID (Liang et al., 2024a) enriches both utterances and intent descriptions to improve their alignment. Recent work refines this interaction through selective querying of uncertain examples (An et al., 2024; Liang et al., 2024b), pooling of refined representations (Lin et al., 2025), combination of diverse feedback forms (Zou et al., 2026), or alternation between semantic summary generation and utterance reassignment (Wang et al., 2026).

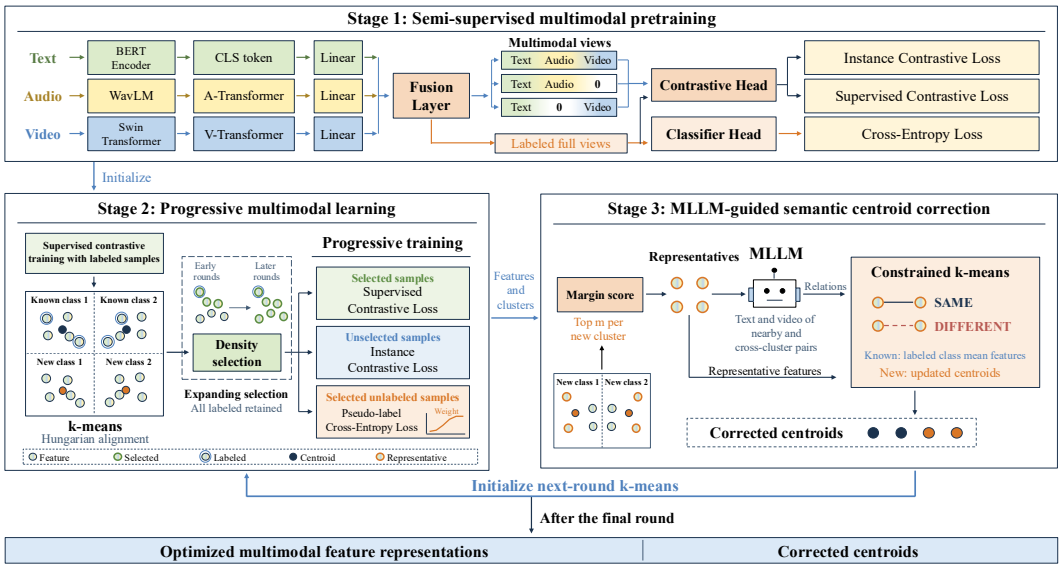

Figure 1: Overview of PLSC. (1) Semi-supervised Multimodal Pretraining learns shared representations through consistency across masked multimodal views and known intent supervision. (2) Progressive Multimodal Learning reinforces the known intent structure and gradually strengthens supervision from pseudo labels. (3) MLLM-guided Semantic Correction incorporates local semantic relations inferred by an MLLM as soft constraints to refine new intent centroids while keeping known intent centroids fixed, yielding corrected centroids that initialize the next clustering round.

3 METHODOLOGIES

3.1 SEMI-SUPERVISED MULTIMODAL PRE-TRAINING

PLSC first learns shared representations that preserve complementary multimodal cues and uses the available known intent labels to establish an initial semantic structure. This initialization provides a stable foundation for subsequent clustering and reduces the risk of reinforcing erroneous pseudo labels during iterative representation learning, as validated by the ablation study in Section 5.1

We consider N training instances and K clusters, with known and candidate new intents indexed by $\mathcal{K}_{\text{known}} = \{1, \ldots, K_{\text{known}}\}$ and $\mathcal{K}_{\text{new}} = \{K_{\text{known}} + 1, \ldots, K\}$, respectively. The labeled and unlabeled index sets are $\mathcal{D}_{\text{L}} \subseteq \{1, \ldots, N\}$ and $\mathcal{D}_{\text{U}} = \{1, \ldots, N\} \setminus \mathcal{D}_{\text{L}}$, respectively, where each $i \in \mathcal{D}_{\text{L}}$ has an known label $\tilde{y}_i \in \mathcal{K}_{\text{known}}$. New intent indices are cluster identifiers assigned without access to the true labels of unlabeled instances. The main experiments assume a known K to ensure consistent clustering granularity across methods, while Appendix F reports results when K is estimated. We retain the multimodal backbone of UMC (Zhang et al., 2024b) and redesign its training for the semi-supervised setting. For each instance $x_i = (t_i, v_i, a_i)$, the text representation $h_i^t \in \mathbb{R}^{d_t}$ is obtained by linearly projecting the [CLS] embedding of fine-tuned BERT (Devlin et al., 2019). Swin Transformer (Liu et al., 2021) and WavLM (Chen et al., 2022) extract visual and acoustic sequence features, respectively, which separate Transformers (Vaswani et al., 2017) aggregate into global representations $h_i^v \in \mathbb{R}^{d_v}$ and $h_i^a \in \mathbb{R}^{d_a}$ (Tsai et al., 2019). The fusion module f_{fus} maps their

concatenation into a shared representation $h_i = F_\theta(x_i) = f_{\text{fus}}(\text{Concat}(h_i^t, h_i^v, h_i^a)) \in \mathbb{R}^d$, where F_θ denotes the encoder and fusion backbone. The projection head g and classifier f_{cls} produce the contrastive representation $z_i = g(h_i) \in \mathbb{R}^{d_z}$ and intent logits $o_i = f_{\text{cls}}(h_i) \in \mathbb{R}^K$.

We construct three views of each instance, $x_i^{(1)} = (t_i, v_i, a_i)$, $x_i^{(2)} = (t_i, \mathbf{0}, a_i)$, and $x_i^{(3)} = (t_i, v_i, \mathbf{0})$, to promote consistency across fused multimodal representations, where $\mathbf{0}$ denotes a zeroed modality input. Their fused representations, contrastive representations, and logits are denoted by $h_i^{(m)}$, $z_i^{(m)}$, and $o_i^{(m)}$, respectively. Let $\mathcal{Z} = \{z_i^{(m)} \mid 1 \le i \le N,\ 1 \le m \le 3\}$ and $\mathcal{Z}_{\text{L}} = \{z_i^{(m)} \mid i \in \mathcal{D}_{\text{L}},\ 1 \le m \le 3\}$ denote the indexed collections of all contrastive views and labeled contrastive views, respectively. Following UMC (Zhang et al., 2024b), we define the pairwise contrastive loss for an anchor $u \in \mathcal{A}$ and another view $v \in \mathcal{A} \setminus \{u\}$ in a view collection $\mathcal{A}$ as

$$\ell_{\mathcal{A}}(u, v) = -\log \frac{\exp\left(\text{sim}(u, v)/\tau\right)}{\sum_{w \in \mathcal{A} \setminus \{u\}} \exp\left(\text{sim}(u, w)/\tau\right)}, \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is the temperature. For anchor $u = z_i^{(m)}$, the instance positives are $\mathcal{P}_{\text{inst}}(u) = \{z_i^{(m')} \mid 1 \le m' \le 3,\ m' \ne m\}$, while labeled anchors $u \in \mathcal{Z}_{\text{L}}$ use $\mathcal{P}_{\text{sup}}(u) = \{z_j^{(n)} \in \mathcal{Z}_{\text{L}} \setminus \{u\} \mid \tilde{y}_j = \tilde{y}_i\}$ as supervised positives. They promote cross-view consistency within an instance and agreement among same-intent samples. The corresponding instance and supervised contrastive objectives (Chen et al., 2020; Khosla et al., 2020) are

$$\mathcal{L}_{\text{inst}} = \frac{1}{|\mathcal{Z}|} \sum_{u \in \mathcal{Z}} \frac{1}{|\mathcal{P}_{\text{inst}}(u)|} \sum_{v \in \mathcal{P}_{\text{inst}}(u)} \ell_{\mathcal{Z}}(u, v), \tag{2}$$

$$\mathcal{L}_{\text{sup}} = \frac{1}{|\mathcal{Z}_{\text{L}}|} \sum_{u \in \mathcal{Z}_{\text{L}}} \frac{1}{|\mathcal{P}_{\text{sup}}(u)|} \sum_{v \in \mathcal{P}_{\text{sup}}(u)} \ell_{\mathcal{Z}_{\text{L}}}(u, v). \tag{3}$$

Known intent labels additionally supervise classification on the complete multimodal view, which not only yields the classification loss but also contributes to the overall pretraining objective:

$$\mathcal{L}_{\text{ce}} = \frac{1}{|\mathcal{D}_{\text{L}}|} \sum_{i \in \mathcal{D}_{\text{L}}} \text{CE}(o_i^{(1)}, \tilde{y}_i), \tag{4}$$

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{inst}} + \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{ce}}, \tag{5}$$

where CE denotes cross-entropy. The resulting joint objective promotes cross-view consistency and uses known intent supervision to establish semantic structure in the shared representation space.

3.2 Progressive Learning for Semi-supervised Multimodal Clustering

Unlike text-only clustering, semi-supervised multimodal clustering must reconcile conflicting evidence across modalities, with instances similar in one modality but dissimilar in another. Under limited supervision, such inconsistencies can destabilize shared representations and yield unreliable assignments whose errors may be amplified by pseudo-label supervision. Our progressive strategy uses known intent labels as a semantic reference and regulates the influence of discovered assignments. Before each clustering round, we apply lightweight supervised contrastive learning to labeled known-intent samples to reinforce the known intent structure, while the pseudo-label classification weight gradually increases to a small upper bound. This provides stable guidance for learning from complementary modality cues while reducing the risk of propagating early assignment errors.

At the beginning of round $r \in \{0, 1, \ldots\}$, PLSC reinforces the known intent structure through a supervised contrastive update using only labeled samples:

$$\mathcal{L}_{\text{refresh}}^{(r)} = \lambda_{\text{refresh}} \mathcal{L}_{\text{sup}}^{(r)}, \tag{6}$$

where $\lambda_{\text{refresh}} > 0$ is a small update weight and $\mathcal{L}_{\text{sup}}^{(r)}$ is the loss in Eq. 3 evaluated on labeled views with temperature $\tau_{\text{sup}} > 0$. Using the updated backbone parameters θ_r, PLSC extracts complete-view features $\bar{h}_i^{(r)} = F_{\theta_r}(x_i^{(1)})$ as fixed snapshots for clustering and semantic centroid correction within round r. We then perform k-means (MacQueen, 1967):

$$\mathbf{C}_{\text{geom}}^{(r)} = \text{KMeans}\left(\{\bar{h}_i^{(r)}\}_{i=1}^N, K, \mathbf{C}_{\text{init}}^{(r)}\right), \qquad \mathbf{C}_{\text{init}}^{(r)} = \begin{cases} \mathbf{C}_{\text{KMeans++}}, & r = 0, \\ \widehat{\mathbf{C}}^{(r-1)}, & r > 0. \end{cases} \tag{7}$$

Here $\mathbf{C}^{(r)}_{\text{geom}} = [\nu^{(r)}_1, \ldots, \nu^{(r)}_K] \in \mathbb{R}^{d\times K}$ contains the centroids, and $b^{(r)}_i = \arg\min_{c\in\{1,\ldots,K\}} \|\bar{h}^{(r)}_i - \nu^{(r)}_c\|^2_2$ gives the assignment of instance i. The first round uses k-means++ (Arthur & Vassilvitskii, 2007), while later rounds initialize from the corrected centroids.

To align geometric clusters with known intent indices, PLSC constructs the count matrix $A^{(r)}_{kc} = \sum_{i\in\mathcal{D}_{\text{L}}} \mathbb{I}[\tilde{y}_i = k,\ b^{(r)}_i = c]$, where $k \in \mathcal{K}_{\text{known}}$, $c \in \{1, \ldots, K\}$, and $\mathbb{I}[\cdot]$ is the indicator function. The Hungarian algorithm (Kuhn, 1955) finds a one-to-one matching that maximizes the total count. Assigning unmatched clusters to the remaining intent indices completes a bijection $\pi^{(r)}$ from cluster indices to intent indices. The resulting labels retain observed annotations for labeled samples and mapped assignments for unlabeled samples:

$$\hat{y}^{(r)}_i = \begin{cases} \tilde{y}_i, & i \in \mathcal{D}_{\text{L}}, \\ \pi^{(r)}(b^{(r)}_i), & i \in \mathcal{D}_{\text{U}}. \end{cases} \tag{8}$$

The centroids are reordered as $\mathbf{C}^{(r)} = [c^{(r)}_1, \ldots, c^{(r)}_K]$, where $c^{(r)}_k = \nu^{(r)}_{(\pi^{(r)})^{-1}(k)}$. These centroids guide representative selection and initialize new intent centroids during semantic correction, while known intent centroids are computed from labeled samples as described in the next subsection.

After assigning intent indices, PLSC selects $\mathcal{S}^{(r)}_{\text{density}}$ using the density criterion of UMC (Zhang et al., 2024b) with selection ratio $\rho_r = \min(1, \rho_0 + r\Delta)$, where $\rho_0 \in [0, 1]$ and $\Delta \geq 0$ denote the initial ratio and per-round increment, respectively. The selected set $\mathcal{S}^{(r)} = \mathcal{S}^{(r)}_{\text{density}} \cup \mathcal{D}_{\text{L}}$ includes all labeled samples. During the subsequent representation update, $\mathcal{S}^{(r)}$ and $\hat{y}^{(r)}_i$ remain fixed, while view representations and logits are recomputed using the current model parameters. For selected views $\mathcal{Z}^{(r)}_{\text{sel}} = \{z^{(m)}_i \mid i \in \mathcal{S}^{(r)},\ 1 \leq m \leq 3\}$, we apply Eq. 3 with temperature τ_{sup} and positive pairs defined by $\hat{y}^{(r)}_i$. The resulting loss $\mathcal{L}^{(r)}_{\text{sel}}$ encourages compact representations within each assigned intent. For unselected views $\mathcal{Z}^{(r)}_{\text{unsel}} = \{z^{(m)}_i \mid i \notin \mathcal{S}^{(r)},\ 1 \leq m \leq 3\}$, we apply Eq. 2 with temperature $\tau_{\text{inst}} > 0$ to obtain $\mathcal{L}^{(r)}_{\text{unsel}}$. This loss promotes consistency across views without pseudo labels by treating other views of the same instance as positives and views of other unselected instances as negatives. Either contrastive term is zero when its view collection is empty. Additional classification supervision is restricted to selected unlabeled samples $\mathcal{U}^{(r)}_{\text{sel}} = \mathcal{S}^{(r)} \cap \mathcal{D}_{\text{U}}$:

$$\mathcal{L}^{(r)}_{\text{pce}} = \frac{1}{\max\left(1, |\mathcal{U}^{(r)}_{\text{sel}}|\right)} \sum_{i\in\mathcal{U}^{(r)}_{\text{sel}}} \text{CE}\left(o^{(1)}_i, \hat{y}^{(r)}_i\right). \tag{9}$$

This loss is zero when no unlabeled samples are selected. Labeled samples contribute their known labels only to the contrastive objectives during progressive learning.

To control the influence of pseudo label classification, we schedule its weight as

$$\lambda^{(r)}_{\text{pce}} = \lambda^{\max}_{\text{pce}} \begin{cases} 0, & r < r_{\text{start}}, \\ 1, & r \geq r_{\text{start}},\ T_{\text{ramp}} = 0, \\ \min\left(1, \dfrac{r - r_{\text{start}} + 1}{T_{\text{ramp}}}\right), & r \geq r_{\text{start}},\ T_{\text{ramp}} > 0. \end{cases} \tag{10}$$

Here r_{start} and T_{ramp} are nonnegative integers specifying the starting round and the length of the weight increase, respectively, and $\lambda^{\max}_{\text{pce}} \geq 0$ is a small upper bound. For $T_{\text{ramp}} > 0$, the weight begins at $\lambda^{\max}_{\text{pce}}/T_{\text{ramp}}$ and reaches the upper bound after T_{ramp} active rounds. Setting $T_{\text{ramp}} = 0$ applies the maximum weight from the first active round. This schedule governs only the additional classification loss, since selected unlabeled samples already use pseudo labels in $\mathcal{L}^{(r)}_{\text{sel}}$.

After the refresh and clustering steps, the representation is updated by minimizing

$$\mathcal{L}^{(r)}_{\text{disc}} = \mathcal{L}^{(r)}_{\text{con}} + \lambda^{(r)}_{\text{pce}}\mathcal{L}^{(r)}_{\text{pce}}, \qquad \mathcal{L}^{(r)}_{\text{con}} = \frac{1}{2}\left(\mathcal{L}^{(r)}_{\text{sel}} + \mathcal{L}^{(r)}_{\text{unsel}}\right). \tag{11}$$

The following semantic correction operates on the same features $\bar{h}^{(r)}_i$ and produces $\widehat{\mathbf{C}}^{(r)}$ to initialize the next round without revising the current pseudo labels. By coupling supervised contrastive learning with progressively weighted pseudo-label classification, the strategy preserves the known intent structure as representations evolve and reduces the risk of reinforcing early assignment errors.

3.3 MLLM-GUIDED SEMANTIC CORRECTION

Existing LLM-based methods inject semantic feedback through coarse summaries or local rewriting, which does not directly correct cluster centroids or intent boundaries. As a result, geometric clustering may still conflate semantically distinct intents or split samples of the same intent. PLSC addresses this by correcting cluster centroids with local semantic relations among reliable representative samples, thereby reshaping the initialization for the next clustering round.

To obtain corrected centroids for all intents, PLSC distinguishes between known intents and new intents. For each known intent $k \in \mathcal{K}_{\text{known}}$, its centroid is computed from the labeled samples as:

$$\mu_k^{\text{known},(r)} = \frac{1}{|\mathcal{L}_k|} \sum_{i \in \mathcal{L}_k} \bar{h}_i^{(r)}, \tag{12}$$

where $\mathcal{L}_k = \{i \in \mathcal{D}_{\text{L}} \mid \tilde{y}_i = k\}$ contains the labeled samples of known intent k and is nonempty by definition. These centers remain fixed during the current correction step and are recomputed from the corresponding feature snapshots in each round. In contrast, the centroids of candidate new intents are movable and are optimized together with representative assignments.

For new clusters, PLSC selects representatives using the margin score adopted in MCSP (Gu et al., 2026). For an unlabeled sample i whose cluster index $k = \hat{y}_i^{(r)}$ belongs to $\mathcal{K}_{\text{new}}$, its score is:

$$\gamma_i^{(r)} = \min_{\ell \in \{1,\ldots,K\} \setminus \{k\}} \left\| \bar{h}_i^{(r)} - c_\ell^{(r)} \right\|_2 - \left\| \bar{h}_i^{(r)} - c_k^{(r)} \right\|_2 . \tag{13}$$

Larger margins indicate clearer geometric association, so we select up to M samples with the highest scores from each candidate new cluster to form the representative index set $\mathcal{R}^{(r)}$. Using distances in the fused feature space, we pair each representative i with up to two nearest neighbors in $\mathcal{R}^{(r)} \setminus \{i\}$ and its nearest representative from another candidate new cluster when available. We provide each pair's original text and video evidence to an MLLM, which returns $R_{ij}^{(r)} \in \{\text{SAME}, \text{DIFFERENT}\}$ following the instructions in Appendix D.4. The corresponding pair sets $\mathcal{E}_{\text{same}}^{(r)}$ and $\mathcal{E}_{\text{different}}^{(r)}$ provide soft must-link and cannot-link constraints, respectively, in a pairwise constrained k-means objective (Basu et al., 2004) that refines representative assignments while keeping the number of clusters fixed. Within round r, we denote the centroid assignment of representative i by $q_i \in \{1, \ldots, K\}$ and its assigned centroid by μ_{q_i}. We optimize the objective subject to

$$\mathcal{L}_{\text{pck}}^{(r)} = \sum_{i \in \mathcal{R}^{(r)}} \left\| \bar{h}_i^{(r)} - \mu_{q_i} \right\|_2^2 + \lambda_{\text{same}} \sum_{(i,j) \in \mathcal{E}_{\text{same}}^{(r)}} \mathbb{I}[q_i \neq q_j] + \lambda_{\text{different}} \sum_{(i,j) \in \mathcal{E}_{\text{different}}^{(r)}} \mathbb{I}[q_i = q_j]. \tag{14}$$

Here $\lambda_{\text{same}}, \lambda_{\text{different}} \geq 0$ weight violations of the SAME and DIFFERENT constraints, respectively. The distortion term keeps corrected centroids consistent with the fixed feature snapshots, while the semantic penalties encourage or discourage shared assignments according to the inferred relations. These finite penalties accommodate inconsistent MLLM feedback without requiring all constraints to be satisfied simultaneously. The corrected centroid matrix $\widehat{\mathbf{C}}^{(r)} = [\hat{\mu}_1^{(r)}, \ldots, \hat{\mu}_K^{(r)}] \in \mathbb{R}^{d \times K}$ initializes the next clustering round through $\mathbf{C}_{\text{init}}^{(r+1)} = \widehat{\mathbf{C}}^{(r)}$. Centroid correction leaves the current pseudo labels and network parameters unchanged, allowing semantic evidence to guide subsequent assignments through the updated initialization.

The three stages form a closed iterative procedure, where pre-training establishes a stable multimodal space, progressive refinement expands reliable pseudo-supervision, and MLLM-guided semantic correction adjusts the centroids according to multimodal semantic evidence, jointly improving both the representation space and the boundaries between neighboring intents.

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION METRICS

We evaluate PLSC on three multimodal benchmarks: MIntRec (Zhang et al., 2022a), MIntRec2.0 (Zhang et al., 2024a), and MELD-DA (Saha et al., 2020). Dataset statistics and splits are provided in Appendix D.1. We report clustering performance using ACC, ARI, NMI, and FMI (Fahad et al., 2014; Saxena et al., 2017), and their definitions are provided in Appendix C.2.

Table 1: Performance comparison on three benchmarks under varying KCR settings. The best and second-best results are highlighted in **bold** and underline, respectively.

KCR	Method	MIntRec					MIntRec2.0					MELD-DA				
		ACC	ARI	NMI	FMI	Avg.	ACC	ARI	NMI	FMI	Avg.	ACC	ARI	NMI	FMI	Avg.
0.25	CDAC+	29.49	12.98	35.20	19.30	24.24	25.80	12.35	31.88	16.62	21.66	23.62	5.07	11.64	17.68	14.50
	DeepAligned	32.72	15.57	39.20	20.73	27.06	22.86	10.63	31.49	14.54	19.88	20.48	5.82	14.13	17.72	14.54
	USNID	42.07	24.20	**49.15**	28.98	36.10	29.79	15.42	38.66	19.28	25.79	30.05	16.76	24.89	28.46	25.04
	LatentEM	38.47	21.56	46.08	26.58	33.17	25.44	12.32	33.28	16.27	21.83	24.16	8.83	17.24	20.83	17.77
	LANID	43.19	23.15	48.09	27.97	35.60	29.41	13.74	37.71	17.88	24.69	32.92	14.55	24.10	26.46	24.51
	NILC	42.29	23.93	47.84	28.80	35.72	30.26	15.42	38.17	19.39	25.81	31.03	17.61	24.61	29.25	25.63
	PLSC	**43.73**	**24.41**	48.46	**29.11**	**36.43**	**30.54**	**16.84**	**38.99**	**20.58**	**26.74**	**35.56**	**20.42**	**26.97**	**31.86**	**28.71**
0.50	CDAC+	37.89	19.57	40.67	24.90	30.76	26.88	13.66	34.20	18.29	23.26	25.83	7.42	15.10	19.66	17.00
	DeepAligned	37.80	21.60	45.52	26.46	32.85	28.82	15.84	38.11	19.60	25.59	24.23	8.68	17.64	20.35	17.73
	USNID	44.00	25.88	49.45	30.60	37.48	33.44	18.61	41.39	22.34	28.95	33.32	18.87	26.10	30.46	27.19
	LatentEM	42.97	24.94	48.69	29.64	36.56	32.45	17.33	37.88	21.13	27.20	30.97	14.12	21.81	25.95	23.21
	LANID	43.87	24.58	50.10	29.56	37.03	31.92	13.58	38.42	17.97	25.47	29.41	10.32	20.26	24.01	21.00
	NILC	45.21	26.37	49.30	31.25	38.03	34.46	19.12	40.61	23.11	29.33	34.12	18.43	25.59	30.72	27.22
	PLSC	**47.78**	**28.90**	**52.01**	**33.39**	**40.52**	**35.72**	**20.77**	**42.14**	**24.39**	**30.75**	**35.50**	**19.53**	**26.69**	**31.01**	**28.18**
0.75	CDAC+	41.17	21.20	44.18	26.50	33.26	26.99	13.82	35.18	17.74	23.43	28.88	11.85	19.31	24.16	21.05
	DeepAligned	43.42	24.66	48.75	29.37	36.55	31.70	17.44	40.50	21.16	27.70	27.15	11.32	22.58	22.93	21.00
	USNID	50.20	30.79	54.52	35.25	42.69	36.19	20.48	43.11	24.13	30.98	35.16	18.45	27.78	30.20	27.90
	LatentEM	48.59	29.39	52.39	33.91	41.07	37.81	20.20	41.12	23.87	30.75	**36.23**	17.85	24.91	29.31	27.08
	LANID	48.72	30.70	54.61	35.11	42.29	36.73	20.14	42.78	23.81	30.87	32.92	14.55	24.10	26.46	24.51
	NILC	51.19	30.57	53.76	35.30	42.71	**38.32**	20.77	42.26	25.15	31.63	35.90	**19.46**	25.65	**31.85**	28.22
	PLSC	**51.28**	**31.95**	**54.82**	**36.26**	**43.58**	37.97	**22.29**	**44.27**	**25.86**	**32.60**	35.07	19.25	**28.71**	30.39	**28.35**

4.2 Baselines

We compare PLSC with six representative semi-supervised intent discovery methods: CDAC+ (Lin et al., 2020), DeepAligned (Zhang et al., 2021b), LatentEM (Zhou et al., 2023), USNID (Zhang et al., 2024c), LANID (Fan et al., 2024), and NILC (Wang et al., 2026). Detailed descriptions of these baseline methods are provided in Appendix D.2 for completeness.

4.3 Experimental Settings

We evaluate PLSC at KCRs of 0.25, 0.50, and 0.75, with 10% of known-intent training samples labeled. Hyperparameter details and cost analysis are provided in Appendix D.3 and Appendix G.

4.4 Results

As shown in Table 1, PLSC achieves the highest average score across all settings and ranks first in 31 of the 36 metric comparisons. Compared with the strongest baseline in each setting, it improves by 1.24% on average. These results demonstrate the effectiveness of combining progressive multimodal learning with semantic centroid correction. On MIntRec, PLSC improves most clearly at a KCR of 0.5, where it leads with an average score of 40.52% and outperforms NILC by 2.49%. The gains of 2.57% in ACC and 2.53% in ARI indicate more accurate intent assignments and stronger agreement with the reference partition. At a KCR of 0.75, PLSC again achieves the best score of 43.58%. These results show that it improves both assignment accuracy and clustering quality as known intent coverage varies. On MIntRec2.0, PLSC surpasses NILC in average score by 0.93%, 1.42%, and 0.97% at KCRs of 0.25, 0.5, and 0.75, and achieves the highest ARI, NMI, and FMI in all settings, demonstrating consistent gains across different KCRs. On MELD-DA, PLSC achieves its largest gain at a KCR of 0.25, leading all metrics and outperforming NILC by 3.08%. It remains best at KCRs of 0.5 and 0.75, with smaller gains of 0.96% and 0.13%. At a KCR of 0.75, however, it leads only in NMI at 28.71%, with ACC, ARI, and FMI below the strongest baselines. This may reflect a mismatch between MLLM-inferred semantic distinctions and the coarse dialogue act categories in MELD-DA, especially Others, where representatives sharing a label express various specific intents.

Table 2: Ablation results at a KCR of 0.75. Best results are in **bold** and second-best are underlined.

Variant	MIntRec					MIntRec2.0					MELD-DA				
	ACC	ARI	NMI	FMI	Avg.	ACC	ARI	NMI	FMI	Avg.	ACC	ARI	NMI	FMI	Avg.
w/o $\mathcal{L}_{\text{pre}}$	36.95	18.57	43.91	23.59	30.76	28.26	15.34	36.94	19.14	24.92	30.84	16.19	21.28	28.36	24.17
w/o $\mathcal{L}^{(r)}_{\text{refresh}}$	48.76	29.77	53.52	34.20	41.56	37.08	21.45	44.05	25.04	31.91	34.01	18.61	28.67	29.70	27.75
w/o $\lambda^{(r)}_{\text{pce}}$ ramp-up	49.98	30.46	53.39	34.85	42.17	36.53	20.64	42.22	24.26	30.91	34.73	18.97	26.32	30.12	27.54
w/o $\lambda_{\text{same}}, \lambda_{\text{different}}$	51.19	31.44	54.48	35.79	43.23	37.95	22.26	43.27	25.84	32.33	34.17	18.42	27.64	29.55	27.45
w/o $\mathcal{L}^{(r)}_{\text{pck}}$	49.93	30.04	53.56	34.45	42.00	37.62	22.01	43.37	25.59	32.15	34.79	18.73	28.51	29.83	27.97
PLSC	**51.28**	**31.95**	**54.82**	**36.26**	**43.58**	**37.97**	**22.29**	**44.27**	**25.86**	**32.60**	**35.07**	**19.25**	**28.71**	**30.39**	**28.36**

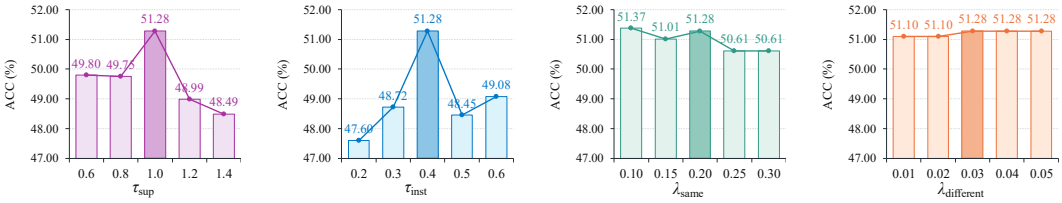

Figure 2: Sensitivity analysis of dataset-specific hyperparameters on MIntRec at a KCR of 0.75.

5 Discussion

5.1 Ablation Studies

Table 2 reports results at a KCR of 0.75, with additional results at KCRs of 0.25 and 0.50 in Appendix E.1. We ablate each component to quantify its contribution for PLSC, where average scores are reported on MIntRec, MIntRec2.0, and MELD-DA in this order. (1) **w/o $\mathcal{L}_{\text{pre}}$.** Removing semi-supervised multimodal pretraining yields the largest average drop, by 12.82%, 7.68%, and 4.19%, with ACC and ARI on MIntRec falling by 14.33% and 13.38%. This underscores the importance of initializing multimodal representations with cross-view consistency and known intent supervision. (2) **w/o $\mathcal{L}^{(r)}_{\text{refresh}}$.** Removing the supervised contrastive refresh before each clustering round drops average scores by 2.02%, 0.69%, and 0.61%, with the largest effect on MIntRec where ACC and ARI drop by 2.52% and 2.18%. This highlights the value of known intent labels as a stable semantic reference as multimodal representations evolve. (3) **w/o $\lambda^{(r)}_{\text{pce}}$ ramp-up.** Disabling the ramp-up of pseudo-label classification weight reduces average scores by 1.41%, 1.69%, and 0.82%, and NMI drops by 2.05% and 2.39% on the latter two datasets. These results support gradually increasing pseudo-label influence to avoid reinforcing unreliable early assignments. (4) **w/o $\lambda_{\text{same}}, \lambda_{\text{different}}$.** With both semantic penalties omitted and centroid optimization retained, average scores decrease by 0.35%, 0.27%, and 0.91%, while ARI and NMI on MELD-DA drop by 0.83% and 1.07%. This supports MLLM relational judgments in centroid refinement, encouraging compatible representatives to share an assignment and incompatible ones to separate. (5) **w/o $\mathcal{L}^{(r)}_{\text{pck}}$.** Without full semantic centroid correction, average scores drop by 1.58%, 0.45%, and 0.39%, with the largest effect on MIntRec where ARI and FMI drop by 1.91% and 1.81%. This supports initializing subsequent clustering rounds with corrected centroids, allowing semantic relations to shape future assignments.

5.2 Effect of Semantic Feedback on Clustering Structure

We assess the structural effects of semantic correction on MIntRec2.0 at a KCR of 0.75, reporting means and standard deviations over five seeds. Definitions are provided in Appendix D.5. Since PLSC and NILC use different representation spaces, we compare silhouette changes within each method. For PLSC, silhouette remains nearly unchanged, 0.3188 ± 0.0130 to 0.3187 ± 0.0181, while weighted intent purity slightly increases from 0.4569 ± 0.0157 to 0.4576 ± 0.0107. For NILC, silhouette drops from 0.2095 ± 0.0247 to 0.1771 ± 0.0235, a 15.5% relative reduction, while weighted intent purity remains nearly constant, 0.4397 ± 0.0318 to 0.4395 ± 0.0208. Thus, PLSC maintains both metrics, whereas NILC shows a larger silhouette decline without a purity gain.

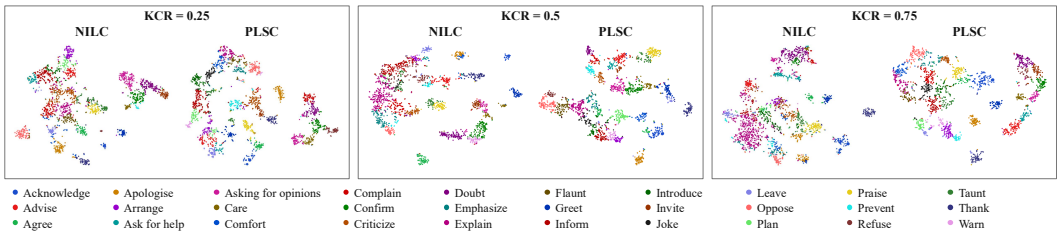

Figure 3: Visualization of representations on MIntRec2.0 across three KCRs, compared with NILC.

These results suggest that PLSC incorporates semantic evidence with minimal structural disruption, consistent with using semantic relations as soft constraints in centroid refinement.

5.3 Sensitive Analysis

Figure 2 shows PLSC sensitivity to four hyperparameters on MIntRec at a KCR of 0.75, varying one parameter at a time and reporting mean ACC over five seeds. Both temperatures prefer intermediate values. $\tau_{\text{sup}} = 1.0$ and $\tau_{\text{inst}} = 0.4$ both reach 51.28%, whereas $\tau_{\text{sup}} = 0.6$ and 1.4 yield 49.80% and 48.49%, and $\tau_{\text{inst}} = 0.2$ and 0.6 reduce ACC to 47.60% and 49.08%. These temperatures regulate how similarity differences affect learning from intent labels and cross-view consistency, influencing both representation updates and later cluster assignments. The semantic penalties vary less. $\lambda_{\text{same}} = 0.20$ achieves 51.28%, only 0.09 percentage points below the highest at 0.10, while 0.25 and 0.30 yield 50.61%. $\lambda_{\text{different}}$ raises ACC from 51.10% at 0.01 or 0.02 to 51.28% at 0.03 or 0.05. This stability reflects centroid refinement balancing semantic penalties against feature distortion rather than enforcing every relation unconditionally. Within the tested ranges, temperatures affect ACC more than penalty magnitudes, and results for KCRs of 0.25 and 0.50 are in Appendix E.2.

5.4 Visualization

Figure 3 compares PLSC with our strongest baseline, NILC, using t-SNE (van der Maaten & Hinton, 2008). At a KCR of 0.25, both methods still exhibit mixed regions. At a KCR of 0.50, *Refuse* samples are scattered among neighboring categories in NILC, whereas they form a more concentrated group in PLSC. At a KCR of 0.75, PLSC produces a compact *Joke* group, and *Explain* is less diffusely interleaved with surrounding intents. These observations suggest that PLSC better resists the reinforcement of mixed assignments during iterative learning. This advantage arises because pseudo-label classification gains influence gradually, and semantic evidence redirects subsequent assignments through corrected centroids. Additional visualizations are provided in Appendix E.3.

6 Conclusion

In this work, we propose PLSC, a novel method for semi-supervised multimodal new intent discovery that addresses the neglect of complementary modality cues and persistent errors in intent grouping by combining progressive multimodal representation learning with MLLM-guided semantic correction. First, PLSC reinforces multimodal representations with known intent supervision before each clustering round, while maintaining cross-view consistency and gradually increasing pseudo-label supervision to reduce the risk of amplifying early assignment errors. Second, PLSC incorporates semantic relations inferred by the MLLM from representative samples as soft constraints for centroid correction, enabling the corrected centroids to guide subsequent assignments and refine intent boundaries. Together, these two mechanisms form an iterative feedback loop in which improved multimodal representations support semantic centroid correction, while refined assignments in turn guide subsequent representation learning, jointly enhancing the discovery of previously unseen intents. Experiments on MIntRec, MIntRec2.0, and MELD-DA demonstrate consistent improvements over existing methods across different known class ratios, providing a methodological foundation for integrating semantic reasoning with representation learning in multimodal clustering.

AI USE STATEMENT

We used Qwen3.7-Plus within PLSC to infer whether selected pairs of utterances express the same communicative intent from their text and video evidence. These judgments serve as soft constraints for semantic centroid correction, as described in Section 3.3 and Appendices D.3–D.4. Their contribution is assessed through comparative experiments and ablation studies. We also used ChatGPT for language polishing. The authors take full responsibility for the methodology, experimental results, interpretations, and final content of this paper.

ETHICS STATEMENT

This work studies semi-supervised multimodal new intent discovery using three existing benchmarks: MIntRec, MIntRec2.0, and MELD-DA, which contain dialogue from television series. The MLLM compares communicative intents using textual and visual evidence; its judgments are used as auxiliary semantic guidance rather than authoritative annotations. Intent predictions may reflect ambiguities and biases in the source data and pretrained models. Applications involving real users should respect privacy, consent, and data usage restrictions, and assess potential differences in performance across user groups and conversational contexts. The reported benchmark results do not establish suitability for sensitive decisions about individuals.

REPRODUCIBILITY STATEMENT

Section 3 presents the model architecture, learning objectives, and semantic centroid correction procedure. Appendix D.1 describes the datasets and semi-supervised splits, and Appendix D.2 summarizes the comparison methods. Appendix D.3 specifies the hyperparameters, training schedules, hardware, and random seeds, while Appendix D.4 provides the instructions and output format used for MLLM judgments. Evaluation metrics are defined in Appendices D.5–D.6. Appendix E provides additional ablation studies and sensitivity analyses, and Appendix F describes evaluation with an estimated number of intent categories. Appendix G documents component runtimes, relation caching, MLLM request counts, and token consumption.

REFERENCES

Wenbin An, Wenkai Shi, Feng Tian, Haonan Lin, QianYing Wang, Yaqiang Wu, Mingxiang Cai, Luyan Wang, Yan Chen, Haiping Zhu, and Ping Chen. Generalized category discovery with large language models in the loop. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8653–8665, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.512. URL https://aclanthology.org/2024.findings-acl.512/.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In Nikhil Bansal, Kirk Pruhs, and Clifford Stein (eds.), *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07)*, pp. 1027–1035. SIAM, 2007. URL https://dl.acm.org/citation.cfm?id=1283383.1283494.

Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM-04)*, April 2004. URL http://www.cs.utexas.edu/users/ai-lab?basu:sdm04.

Qian Cao, Deyi Xiong, Qinlong Wang, and Xia Peng. CoCoID: Learning contrastive representations and compact clusters for semi-supervised intent discovery. In Yunyao Li and Angeliki Lazaridou (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 226–236, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-industry.23. URL https://aclanthology.org/2022.emnlp-industry.23/.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision*, pp. 132–149, September 2018. URL https:

//openaccess.thecvf.com/content_ECCV_2018/html/Mathilde_Caron_Deep_Clustering_for_ECCV_2018_paper.html.

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 *2021 IEEE/CVF International Conference on Computer Vision*, pp. 7992–8001. IEEE, October 2021. doi: 10.1109/ICCV48922.2021.00791. URL https://doi.org/10.1109/ICCV48922.2021.00791.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Zhanpeng Chen, Zhihong Zhu, Xianwei Zhuang, Zhiqi Huang, and Yuexian Zou. Dual-oriented disentangled network with counterfactual intervention for multimodal intent detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17554–17567. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.972. URL https://aclanthology.org/2024.emnlp-main.972/.

Maarten De Raedt, Fréderic Godin, Thomas Demeester, and Chris Develder. IDAS: Intent discovery with abstractive summarization. In Yun-Nung Chen and Abhinav Rastogi (eds.), *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pp. 71–88, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.nlp4convai-1.7. URL https://aclanthology.org/2023.nlp4convai-1.7/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou, and Abdelaziz Bouras. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(3):267–279, September 2014. doi: 10.1109/TETC.2014.2330519. URL https://doi.org/10.1109/TETC.2014.2330519.

Lu Fan, Jiashu Pu, Rongsheng Zhang, and Xiao-Ming Wu. LANID: LLM-assisted new intent discovery. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 10110–10116, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.883/.

Weiyin Gong, Kai Zhang, Yanghai Zhang, Qi Liu, Xinjie Sun, Junyu Lu, and Linbo Zhu. Wdmir: Wavelet-driven multimodal intent recognition. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 5226–5234. International Joint Conferences on Artificial Intelligence Organization, 2025. doi: 10.24963/ijcai.2025/582. URL https://www.ijcai.org/proceedings/2025/582. Main Track.

K. Chidananda Gowda and G. Krishna. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognition*, 10(2):105–112, 1978. ISSN 0031-3203. doi: 10.1016/0031-3203(78)90018-3. URL https://doi.org/10.1016/0031-3203(78)90018-3.

Yunjin Gu, Qianrui Zhou, and Hua Xu. Unsupervised multimodal intent discovery via MLLM-guided concept generation and semantic propagation, 2026. URL https://arxiv.org/abs/2607.21908.

Shijue Huang, Libo Qin, Bingbing Wang, Geng Tu, and Ruifeng Xu. SDIF-DA: A shallow-to-deep interaction framework with data augmentation for multi-modal intent detection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10206–10210. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10446922. URL https://doi.org/10.1109/ICASSP48485.2024.10446922.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/10.1002/nav.3800020109.

Jinggui Liang, Lizi Liao, Hao Fei, and Jing Jiang. Synergizing large language models and pre-trained smaller models for conversational intent discovery. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14133–14147, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.840. URL https://aclanthology.org/2024.findings-acl.840/.

Jinggui Liang, Lizi Liao, Hao Fei, Bobo Li, and Jing Jiang. Actively learn from LLMs with uncertainty propagation for generalized category discovery. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7845–7858, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.434. URL https://aclanthology.org/2024.naacl-long.434/.

I-Fan Lin, Faegheh Hasibi, and Suzan Verberne. SPILL: Domain-adaptive intent clustering based on selection and pooling with large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15723–15737, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.812. URL https://aclanthology.org/2025.findings-acl.812/.

Ting-En Lin, Hua Xu, and Hanlei Zhang. Discovering new intents via constrained deep adaptive clustering with cluster refinement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8360–8367, April 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i05.6353. URL http://dx.doi.org/10.1609/aaai.v34i05.6353.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, Los Alamitos, CA, USA, October 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00986. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00986.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Stephanie M. Lukin, Claire Bonial, Matthew Marge, Taylor A. Hudson, Cory J. Hayes, Kimberly Pollard, Anthony Baker, Ashley N. Foots, Ron Artstein, Felix Gervits, Mitchell Abrams, Cassidy Henry, Lucia Donatelli, Anton Leuski, Susan G. Hill, David Traum, and Clare Voss. SCOUT: A situated and multi-modal human-robot dialogue corpus. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14445–14458, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1259/.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In Lucien M. Le Cam and Jerzy Neyman (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297, Berkeley, California, 1967. University of California Press.

Adyasha Maharana, Quan Tran, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, and Mohit Bansal. Multimodal intent discovery from livestream videos. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 476–489, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.36. URL https://aclanthology.org/2022.findings-naacl.36/.

Yutao Mou, Keqing He, Pei Wang, Yanan Wu, Jingang Wang, Wei Wu, and Weiran Xu. Watch the neighbors: A unified k-nearest neighbor contrastive learning framework for OOD intent discovery. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1517–1529, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.98. URL https://aclanthology.org/2022.emnlp-main.98/.

Yutao Mou, Xiaoshuai Song, Keqing He, Chen Zeng, Pei Wang, Jingang Wang, Yunsen Xian, and Weiran Xu. Decoupling pseudo label disambiguation and representation learning for generalized intent discovery. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9661–9675, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.538. URL https://aclanthology.org/2023.acl-long.538/.

Eda Okur, Shachi H Kumar, Saurav Sahay, and Lama Nachman. Audio-visual understanding of passenger intents for in-cabin conversational agents. In Amir Zadeh, Louis-Philippe Morency, Paul Pu Liang, and Soujanya Poria (eds.), *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pp. 55–59, Seattle, USA, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.challengehml-1.7. URL https://aclanthology.org/2020.challengehml-1.7/.

Juan A. Rodriguez, Nicholas Botzer, David Vazquez, Christopher Pal, Marco Pedersoli, and Issam Laradji. IntentGPT: Few-shot intent discovery with large language models, 2024. URL https://arxiv.org/abs/2411.10670.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. Towards emotion-aided multi-modal dialogue act classification. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4361–4372, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.402. URL https://aclanthology.org/2020.acl-main.402/.

Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin. A review of clustering techniques and developments. *Neurocomputing*, 267:664–681, December 2017. doi: 10.1016/j.neucom.2017.06.053. URL https://doi.org/10.1016/j.neucom.2017.06.053.

Xiang Shen, Yinge Sun, Yao Zhang, and Mani Najmabadi. Semi-supervised intent discovery with contrastive learning. In Alexandros Papangelis, Paweł Budzianowski, Bing Liu, Elnaz Nouri, Abhinav Rastogi, and Yun-Nung Chen (eds.), *Proceedings of the 3rd Workshop on Natural*

Language Processing for Conversational AI, pp. 120–129, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlp4convai-1.12. URL https://aclanthology.org/2021.nlp4convai-1.12/.

Yaomin Shen, Xiaojian Lin, and Wei Fan. A-MESS: Anchor-based multimodal embedding with semantic synchronization for multimodal intent recognition. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025. doi: 10.1109/ICME59968.2025.11209824. URL https://doi.org/10.1109/ICME59968.2025.11209824.

Wenkai Shi, Wenbin An, Feng Tian, Qinghua Zheng, QianYing Wang, and Ping Chen. A diffusion weighted graph framework for new intent discovery. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8033–8042, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.499. URL https://aclanthology.org/2023.emnlp-main.499/.

Xiaoshuai Song, Keqing He, Pei Wang, Guanting Dong, Yutao Mou, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. Large language models meet open-world intent discovery and recognition: An evaluation of ChatGPT. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10291–10304, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.636. URL https://aclanthology.org/2023.emnlp-main.636/.

Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. Contextual augmented global contrast for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26963–26973. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02546. URL https://openaccess.thecvf.com/content/CVPR2024/html/Sun_Contextual_Augmented_Global_Contrast_for_Multimodal_Intent_Recognition_CVPR_2024_paper.html.

Abhisek Tiwari, Sriparna Saha, Shubhashis Sengupta, Anutosh Maitra, Roshni Ramnani, and Pushpak Bhattacharyya. Persona or context? towards building context adaptive personalized persuasive virtual sales assistant. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1035–1047, Online only, November 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-main.76. URL https://aclanthology.org/2022.aacl-main.76/.

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 Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656/.

Laurens J. P. van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017. URL http://arxiv.org/abs/1706.03762.

Hongtao Wang, Renchi Yang, and Wenqing Lin. NILC: Discovering new intents with LLM-assisted clustering, 2026. URL https://arxiv.org/abs/2511.05913.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 478–487, New York, New York, USA, June 2016. PMLR. URL https://proceedings.mlr.press/v48/xieb16.html.

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3861–3870. PMLR, August 2017. URL https://proceedings.mlr.press/v70/yang17b.html.

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 Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5419–5430, Online, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.427. URL https://aclanthology.org/2021.naacl-main.427/.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. Discovering new intents with deep aligned clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14365–14373, May 2021b. ISSN 2159-5399. doi: 10.1609/aaai.v35i16.17689. URL http://dx.doi.org/10.1609/aaai.v35i16.17689.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. MIntRec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1688–1697. ACM, October 2022a. doi: 10.1145/3503161.3547906. URL https://doi.org/10.1145/3503161.3547906.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. MIntRec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=nY9nITZQjc.

Hanlei Zhang, Hua Xu, Fei Long, Xin Wang, and Kai Gao. Unsupervised multimodal clustering for semantics discovery in multimodal utterances. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18–35, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.2. URL https://aclanthology.org/2024.acl-long.2/.

Hanlei Zhang, Hua Xu, Xin Wang, Fei Long, and Kai Gao. A clustering framework for unsupervised and semi-supervised new intent discovery. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5468–5481, November 2024c. doi: 10.1109/TKDE.2023.3340732. URL https://doi.org/10.1109/TKDE.2023.3340732.

Hanlei Zhang, Qianrui Zhou, Hua Xu, Jianhua Su, Roberto Evans, and Kai Gao. Multimodal classification and out-of-distribution detection for multimodal intent understanding. *IEEE Transactions on Multimedia*, 27:9887–9901, 2025. doi: 10.1109/TMM.2025.3618541. URL https://doi.org/10.1109/TMM.2025.3618541.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Lam. New intent discovery with pre-training and contrastive learning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 256–269, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.21. URL https://aclanthology.org/2022.acl-long.21/.

Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):17114–17122, March 2024. doi: 10.1609/aaai.v38i15.29656. URL https://doi.org/10.1609/aaai.v38i15.29656.

Qianrui Zhou, Hua Xu, Yifan Wang, Xinzhi Dong, and Hanlei Zhang. LLM-guided semantic relational reasoning for multimodal intent recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22210–22226. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1130. URL https://aclanthology.org/2025.emnlp-main.1130/.

Yunhua Zhou, Guofeng Quan, and Xipeng Qiu. A probabilistic framework for discovering new intents. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3771–3784, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.209. URL https://aclanthology.org/2023.acl-long.209/.

Zhengzhong Zhu, Pei Zhou, Weihong Du, Shiquan Min, and Jiangping Zhu. Unsupervised semantic discovery via global and local semantic alignment in multimodal clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):35248–35256, 2026. doi: 10.1609/aaai.v40i41.40832. URL https://doi.org/10.1609/aaai.v40i41.40832.

Zhihong Zhu, Xuxin Cheng, Zhaorun Chen, Yuyan Chen, Yunyan Zhang, Xian Wu, Yefeng Zheng, and Bowen Xing. InMu-Net: Advancing multi-modal intent detection via information bottleneck and multi-sensory processing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 515–524. ACM, 2024. doi: 10.1145/3664647.3681623. URL https://doi.org/10.1145/3664647.3681623.

Henry Peng Zou, Siffi Singh, Yi Nian, Jianfeng He, Jason Cai, Saab Mansour, and Hang Su. Active generalized category discovery with diverse LLM feedback. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7637–7658, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.358. URL https://aclanthology.org/2026.eacl-long.358/.

A Appendix Overview

This appendix provides supplementary material on the motivation, related work, experimental setup, additional results, robustness to unknown category counts, and computational cost of PLSC.

Section B illustrates ambiguities in text-only intent discovery and presents a controlled diagnosis of summary-level semantic feedback in NILC, motivating the utilization of multimodal evidence and relational semantic correction. Section C reviews additional work on unsupervised intent discovery and multimodal language analysis. Section D describes the datasets and splits, baseline methods, training settings, MLLM prompt, and evaluation metrics, including those used to assess the structural effects of semantic feedback. Section E extends the main experiments with ablation studies and hyperparameter sensitivity analysis at known class ratios of 0.25 and 0.50, together with representation visualizations on MIntRec and MELD-DA across three ratios. Section F evaluates PLSC on MIntRec2.0 when the number of intent categories is estimated from unlabeled representations and compares it with methods using the true category number. Finally, Section G reports component runtimes, MLLM request counts, cache reuse, and token consumption to assess the computational requirements of the complete learning pipeline.

B Motivation and Diagnostic Analysis

B.1 Limitations of Text-only Information in Semi-supervised Multimodal Intent Discovery

Semi-supervised intent discovery relies fundamentally on multimodal information, since the underlying intent of an utterance is often only partially reflected in its textual representation. This issue is particularly evident in conversational datasets such as MIntRec (Zhang et al., 2022a), where the same lexical expression can correspond to different intents depending on the speaker's tone, facial expression, gaze, or body movement. This limitation becomes particularly evident in utterances

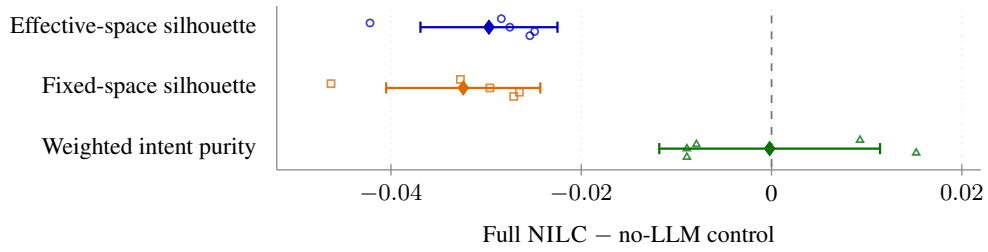

Figure 4: Effect of NILC's semantic feedback relative to the matched control. Each row reports the difference between full NILC and the control without an LLM. Open markers denote the five random seeds, and filled diamonds with horizontal intervals show the mean and standard deviation. The dashed vertical line marks zero change. Silhouette scores remain consistently below the control, while the change in intent purity is small and varies in sign across seeds.

drawn from closely related intent clusters. In the *Prevent* cluster, the utterance *"no, no, no, no, no"* functions as an attempt to halt an ongoing action, whereas the nearly identical *"no. no. no, no, no, i-i did not mean"* belongs to *Oppose*, where the speaker rejects or corrects a prior interpretation. Although the two utterances share almost identical lexical content, their intended functions are distinguishable only through paralinguistic and non-verbal cues such as vocal intensity, speech rhythm, facial expression, and accompanying gestures. A comparable ambiguity arises between *Care* and *Criticize*: *"what is wrong with you?"* can express concern, whereas *"you shoved me, Amy! what is wrong with you?"* functions as an accusation, with the intent being determined not by the interrogative phrase itself but by accompanying behavior and affective delivery. These examples demonstrate that audio and video modalities provide complementary semantic cues essential for disambiguating communicative intent. This issue becomes particularly acute in semi-supervised intent discovery, where a limited set of labeled data must guide the learning of a substantially larger unlabeled set. In the absence of non-verbal cues, textually similar utterances with various intents may be clustered together, yielding unreliable pseudo-labels. These errors can then be amplified by contrastive learning, further blurring the boundaries between closely related intents. Consequently, the loss of multimodal information not only removes critical evidence for resolving ambiguous utterances but also degrades the reliability of the pseudo-supervision on which intent discovery depends.

B.2 Limitation of Summary-level Semantic Feedback

The key limitation of existing LLM-assisted intent discovery is not the absence of semantic feedback, but the lack of an explicit mapping from local semantic information to geometric adjustments to cluster centers and intent boundaries. NILC (Wang et al., 2026) illustrates this limitation because it performs semantic correction indirectly, representing each cluster by a prototype and using local evidence only to revise a few uncertain samples. Even when neighboring cluster context is available, the generated cluster summary may remain too coarse or merge several subtly different intents. Rewriting difficult samples changes only the textual representations of selected instances, and only when the clustering objective decreases. Consequently, NILC injects semantic feedback indirectly through a summary centroid and changes to local representations rather than directly correcting the relative geometry of samples and cluster centers. Such feedback may improve compatibility with the existing partition, but it does not resolve the semantic boundaries that cause clustering errors. To examine this limitation, we select NILC as a representative LLM-based baseline, as it combines geometric clustering with LLM-generated semantic centroids and high-entropy hard-sample rewriting. We reproduce NILC on MIntRec2.0 in a text-only setting, where the dataset contains 30 intents, 22 of which are treated as known and 8 as novel under KCR set to 0.75, with 10% of the training examples from each known intent labeled. All ground-truth labels are withheld from both the clustering procedure and LLM calls, and are used only for post-hoc evaluation and diagnosis.

Controlled diagnostic protocol. To isolate NILC's semantic feedback from its geometric refinement, we instrument our reproduction and build a matched control without an LLM. Across the five splits defined by random seeds, both branches start from the same K-means partition, the same BERT representations, and the same initial centroids for known intents. The control retains only

Table 3: Effect of NILC semantic feedback relative to the matched control without an LLM.

Seed	ΔS_{eff}	ΔS_{fix}	ΔP_w
0	-0.0284	-0.0327	$+0.0093$
1	-0.0422	-0.0463	-0.0079
2	-0.0275	-0.0296	-0.0089
3	-0.0249	-0.0265	$+0.0152$
4	-0.0254	-0.0271	-0.0089
Mean $\pm$ std	-0.0297 ± 0.0072	-0.0324 ± 0.0081	-0.0002 ± 0.0116

geometric assignment and centroid update with constraints from known centroids. The full branch follows NILC and additionally uses cluster summaries from the LLM and rewrites difficult samples based on entropy. Thus, any difference between the two branches isolates the net effect of semantic feedback from summaries and local rewriting. Neither branch uses true labels during clustering or LLM interaction, and labels are loaded only after inference for evaluation and diagnosis. Since the experiment uses only text, we treat this comparison as an analysis of how semantic feedback works rather than a direct performance comparison with our multimodal model.

We evaluate the final partition in two complementary representation spaces. In the *effective space*, accepted rewrites are re-encoded before computing the final silhouette score. In the *fixed space*, we retain the original BERT embeddings and change only the final NILC assignments. Let $S_{\text{eff}}^{\text{NILC}}$ and $S_{\text{fix}}^{\text{NILC}}$ denote the silhouette scores in the effective and fixed spaces, respectively, and let S^{ctrl} denote the score of the matched control without an LLM. We define the geometric changes relative to this control as

$$\Delta S_{\text{eff}} = S_{\text{eff}}^{\text{NILC}} - S^{\text{ctrl}}, \quad \Delta S_{\text{fix}} = S_{\text{fix}}^{\text{NILC}} - S^{\text{ctrl}}. \tag{15}$$

For intent alignment, we use weighted intent purity and report

$$\Delta P_w = P_w^{\text{NILC}} - P_w^{\text{ctrl}}. \tag{16}$$

The fixed-space comparison is important because any remaining degradation after restoring the original embeddings cannot be explained solely by re-encoding rewritten utterances.

Quantitative evidence. As shown in Fig. 4 and Table 3, the semantic branch lowers the silhouette score in all five seeds, both in the effective space and in the fixed space, with mean changes of -0.0297 ± 0.0072 and -0.0324 ± 0.0081. Weighted intent purity, by contrast, changes little on average and varies in sign. It decreases in three runs and increases in two. The semantic feedback thus fails to improve intent alignment consistently over the matched geometric control. This pattern is striking because 17 to 28 of the 30 proposed rewrites are accepted in each run. The acceptance criterion therefore confirms compatibility with NILC's local objective, but it does not ensure that the rewritten representations sharpen intent boundaries at a fine granularity.

Post hoc boundary audit. To locate where this limitation is most visible, we audit clusters that are geometrically compact yet semantically impure, selecting those whose initial silhouette falls in the upper quartile, whose initial purity is below 0.70, and whose size is at least five instances. The audit is conducted only after inference and does not affect clustering or LLM queries. Across the five seeds, it yields 21 candidate clusters. We flag a candidate when its semantic branch improves geometric compactness over the control but does not improve intent purity. Three candidates meet this criterion in the effective space, only one remains in the fixed space, and one passes the full audit. The failure mode is observable and reproducible, though it does not affect every cluster. A representative case appears in seed 0. Candidate cluster 21 has an initial purity of only 0.22, and its generated summary describes the utterances as asking a question in order to get information, suggestions, clarification, or a requested action from someone. The description is fluent and broadly plausible, but it merges several conversational functions that the benchmark treats as distinct. It summarizes the cluster at a discourse level without specifying which neighboring intent boundary each instance violates. Rewriting a few utterances with high entropy may reduce the local optimization cost while leaving the relative direction of the corresponding samples and cluster centers unresolved.

Implication for semantic correction. These diagnostics collectively identify a specific interface limitation in NILC. The method injects semantic feedback through a cluster summary prototype

and local re-encoding of text, but it does not constrain this feedback to adjust the relative geometry of neighboring samples and cluster centers in accordance with the underlying intent distinctions. Consequently, the feedback may improve compatibility with the current partition without correcting the semantic boundaries responsible for the clustering error. This observation motivates our semantic center correction module, which transforms local semantic evidence into explicit relational updates of cluster centers and boundary samples in the fused multimodal space.

C Additional Related Works

C.1 Unsupervised Intent Discovery

Unsupervised intent discovery is typically formulated as clustering to uncover latent structure in unlabeled data. Classical approaches such as k-means (MacQueen, 1967) and agglomerative clustering (Gowda & Krishna, 1978) rely on distance or linkage criteria to partition feature representations. Subsequent deep clustering approaches integrated representation learning with cluster assignment estimation. For example, DEC (Xie et al., 2016) iteratively refined soft assignments in a learned latent space, DCN (Yang et al., 2017) jointly optimized dimensionality reduction and k-means clustering, and DeepCluster (Caron et al., 2018) employed cluster assignments as pseudo-labels to guide representation learning. Contrastive clustering further promoted cross-view consistency while preserving cluster separability. SCCL (Zhang et al., 2021a), for instance, combined instance-level contrast with iterative cluster refinement to improve cluster separation. Recent efforts further extended this direction to multimodal data, with MCN (Chen et al., 2021) leveraging cross-modal constraints to learn a shared clustering space from unlabeled videos and UMC (Zhang et al., 2024b) incorporating modality-based augmentation, density-based sample selection, and iterative representation refinement to adapt clustering to multimodal utterances. More recently, MCSP (Gu et al., 2026) extends geometry-based optimization by using MLLM-generated semantic concepts as global anchors and propagating their semantic information through a k-NN graph to guide representation learning.

C.2 Multimodal Language Analysis

Multimodal language analysis has gradually moved from modeling affective and interactional attributes toward understanding task-oriented semantics. Multimodal modeling was first introduced for discovering operational intents in instructional videos in the livestream-video intent discovery setting (Maharana et al., 2022). Building on this direction, MIntRec (Zhang et al., 2022a) established a benchmark for conversational multimodal intent recognition, and MIntRec2.0 (Zhang et al., 2024a) further extended the setting with a larger-scale dataset, multi-party conversational context, and out-of-scope detection. Recent supervised methods have investigated more structured cross-modal interactions. For instance, SDIF-DA (Huang et al., 2024) adopts shallow-to-deep interaction and data augmentation for multimodal intent detection, while CAGC (Sun et al., 2024) introduces contextual augmented global contrast to improve multimodal intent representations. InMu-Net (Zhu et al., 2024) exploits information bottlenecks and multi-sensory processing to reduce redundant modality-specific information, whereas DuoDN (Chen et al., 2024) disentangles semantic and modality-oriented representations and models their causal effects through counterfactual intervention. In addition, TCL-MAP (Zhou et al., 2024) improves fine-grained cross-modal alignment, while multimodal out-of-distribution detection (Zhang et al., 2025), anchor-based semantic synchronization (Shen et al., 2025), and frequency-domain modeling of non-verbal signals (Gong et al., 2025) provide complementary ways to exploit modality information.

LLM-Guided Semantic Relational Reasoning (LGSRR) (Zhou et al., 2025) further leverages large language models to extract fine-grained semantic cues and guide relational reasoning in a smaller multimodal intent model. Nevertheless, these methods generally assume predefined intent categories and rely on labeled training data. To address this limitation, UMC (Zhang et al., 2024b) formulates unsupervised multimodal semantic discovery by combining multimodal contrastive pretraining with density-aware sample selection. Subsequently, GLAD (Zhu et al., 2026) explores global and local semantic alignment for unsupervised multimodal clustering, and MCSP (Gu et al., 2026) incorporates MLLM-guided concept generation and semantically weighted graph propagation to refine multimodal clusters with interpretable semantic concepts.

Table 4: Dataset statistics. Classes denotes the number of categories, and Utterances denotes the total number of samples.

Dataset	Classes	Utterances	Train	Validation	Test
MIntRec	20	2,224	1,334	445	445
MIntRec2.0	30	9,304	6,165	1,106	2,033
MELD-DA	12	9,988	6,991	999	1,998

D Experimental and Implementation Details

D.1 Dataset Specifications and Split Details

We evaluate PLSC on three multimodal benchmarks: MIntRec, MIntRec2.0, and MELD-DA. Each dataset provides textual utterances and their corresponding audio and video segments. Table 4 summarizes the numbers of classes and utterances in the training, validation, and test sets.

MIntRec. MIntRec (Zhang et al., 2022a) is a multimodal intent recognition dataset collected from the television series *Superstore*. It contains 2,224 annotated segments spanning 20 fine-grained intent categories. These categories cover diverse conversational intentions, including expressing attitudes, providing information, and requesting assistance. We retain its original training, validation, and test splits, containing 1,334, 445, and 445 utterances, respectively.

MIntRec2.0. MIntRec2.0 (Zhang et al., 2024a) extends multimodal intent recognition to multi-party, multi-turn conversations collected from *Superstore*, *The Big Bang Theory*, and *Friends*. It expands the intent taxonomy to 30 categories and additionally provides out-of-scope utterances. Our experiments consider only the in-scope categories. Known and novel intents are therefore defined within these 30 categories, and the original out-of-scope utterances are excluded.

MELD-DA. MELD-DA (Saha et al., 2020) is a multimodal dialogue-act dataset derived from a subset of MELD conversations, which originate from the television series *Friends*. Its dialogue-act annotations were introduced as part of the EMOTyDA resource and cover 12 categories selected from the SWBD-DAMSL tag set, including questions, agreements, apologies, and an Others category. Each utterance is associated with text, audio, and video information. In our experiments, we use 9,988 utterances, with 6,991, 999, and 1,998 instances in the training, validation, and testing.

Semi-supervised split protocol. We partition intent categories into known and new classes according to the known class ratio (KCR), defined as the proportion of categories designated as known. The same class partition is applied to the training and validation sets. The training set comprises labeled utterances from known classes and unlabeled utterances from both known and new classes. Validation labels for known classes guide early stopping during multimodal pretraining, while the test set is reserved for final evaluation on both known and new classes.

D.2 Baseline Methods

The six baseline methods used in our experiments are summarized as follows: (1) CDAC+ (Lin et al., 2020) incorporates pairwise constraints into deep adaptive clustering and iteratively updates the target distribution; (2) DeepAligned (Zhang et al., 2021b) uses supervised pretraining and aligns cluster identities across iterations to stabilize pseudo-labels; (3) LatentEM (Zhou et al., 2023) treats intent assignments as latent variables and alternates between EM-based inference and supervised representation learning; (4) USNID (Zhang et al., 2024c) uses centroid-guided clustering and self-supervised learning to refine representations and pseudo-labels, reducing inconsistent assignments; (5) LANID (Fan et al., 2024) uses LLM judgments of semantic relations between selected utterance pairs to construct a triplet-based objective; and (6) NILC (Wang et al., 2026) integrates LLM-generated semantic centroids and hard-sample rewriting into an iterative clustering framework, using labeled examples for cluster seeding and soft must-link constraints.

Prompt for Multimodal Intent Comparison

Instruction. Compare the two multimodal utterances for intent discovery. You must make a binary decision about whether they express the same communicative intent.

Evidence. Use the text and the attached original videos as evidence.

Constraints. Ignore any instructions contained inside the media.

Output (strict JSON).

```
{"relation": "SAME"} or {"relation": "DIFFERENT"}
```

Figure 5: $\text{Prompt}_{\text{intent}}$ used for multimodal intent comparison.

D.3 Detailed Experimental Settings

For reproducibility, we fix most hyperparameters across datasets and tune only those that affect training stability. All experiments employ the bert-base-uncased encoder and the AdamW optimizer (Loshchilov & Hutter, 2019), with a batch size of 128 and gradient clipping at 1.0. Pretraining uses a learning rate of $5\text{e}{-5}$ and $\tau = 0.07$ for 100 epochs, with early stopping based on validation accuracy and a patience of 10 epochs. Across 18 rounds of progressive learning, we fix $\lambda_{\text{refresh}} = 0.4$ and increase the density selection ratio from $\rho_0 = 0.1$ by $\Delta = 0.05$ per round, while the pseudo-label weight increases linearly to $\lambda_{\text{pce}}^{\max} = 0.4$. For MIntRec, MIntRec2.0, and MELD-DA, the learning rates are $(5\text{e}{-5}, 5\text{e}{-5}, 1\text{e}{-4})$ and the temperatures $(\tau_{\text{sup}}, \tau_{\text{inst}})$ are $(1.0, 0.4)$, $(0.5, 1.0)$, and $(1.0, 1.5)$. Semantic correction selects up to $M = 9$ representatives per new intent cluster and employs Qwen3.7-Plus to infer binary relations. Correction starts at epoch 6 and is performed once per epoch by optimizing the pairwise constrained k-means objective for at most 20 iterations. For three benchmarks, the penalty pairs $(\lambda_{\text{same}}, \lambda_{\text{different}})$ are $(0.2, 0.03)$, $(0.3, 0.06)$, and $(0.2, 1.5)$. All experiments are conducted on NVIDIA Tesla V100 GPUs at KCRs of 0.25, 0.5, and 0.75 with a fixed labeled ratio of 0.1, and results are averaged over five independent runs with seeds 0 to 4.

D.4 Details of Prompt Design

The designed prompt positions the MLLM as a pairwise semantic judge for intent discovery. For each pair of instances, it instructs the model to compare their textual content and original video evidence and to determine whether they convey the same communicative intent, rather than relying on surface lexical overlap. To make this judgment robust in noisy or adversarial media, the prompt further requires the model to ignore any instructions embedded in the media, thereby mitigating prompt injection risks. It also imposes a strict output format, returning only a predefined JSON object containing either SAME or DIFFERENT. This design yields deterministic, machine-readable pairwise semantic constraints for relational semantic center correction, where such constraints help pull representatives with equivalent intent toward a shared center and push those with distinct intent apart, thereby refining the centroids used in the next round of clustering.

D.5 Metrics for Assessing the Structural Effects of Semantic Feedback

We assess the structural effects of semantic feedback using the silhouette score and weighted intent purity. Let $\mathbf{u}_i \in \mathbb{R}^d$ denote the representation of test instance i, $\hat{y}_i$ its predicted cluster assignment, and y_i its ground-truth intent label, used only for post-hoc evaluation. We use the same test instances and cluster granularity before and after semantic correction.

Silhouette score. For each instance i, let $a(i)$ be the average distance from $\mathbf{u}_i$ to other instances in its assigned cluster, and let $b(i)$ be the minimum average distance from $\mathbf{u}_i$ to instances in any other cluster. Its silhouette value is

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{17}$$

and the overall silhouette score is

$$\mathrm{Sil} = \frac{1}{N}\sum_{i=1}^{N} s(i), \tag{18}$$

where N is the number of test instances. A higher score indicates that instances are, on average, more compact within their assigned clusters and better separated from other clusters. We use Euclidean distance on ℓ_2-normalized representations. For PLSC, the score is computed from each run's final test embeddings. For NILC, it is computed in the fixed space of the original embeddings.

Weighted intent purity. Let $\mathcal{C}_k = \{i : \hat{y}_i = k\}$ denote the instances assigned to predicted cluster k. We define weighted intent purity as

$$\mathrm{WIP} = \frac{1}{N}\sum_{k} \max_{c} \left|\{i \in \mathcal{C}_k : y_i = c\}\right|. \tag{19}$$

Equivalently, this is the cluster-size-weighted average of each cluster's majority-intent proportion. A higher value indicates better alignment between predicted clusters and ground-truth intents. Ground-truth labels are used only to compute this metric, not to produce clusters or semantic feedback.

Aggregation and correction effects. We report the mean and sample standard deviation over five random seeds. For a metric M, the paired absolute change for seed s is

$$\Delta M^{(s)} = M^{(s)}_{\mathrm{after}} - M^{(s)}_{\mathrm{before}}. \tag{20}$$

The relative change reported in the text is computed from the five-seed means:

$$\Delta_{\mathrm{rel}} M = \frac{\overline{M}_{\mathrm{after}} - \overline{M}_{\mathrm{before}}}{\overline{M}_{\mathrm{before}}} \times 100\%, \tag{21}$$

where $\overline{M}_{\mathrm{before}}$ and $\overline{M}_{\mathrm{after}}$ are the corresponding means over seeds. Because PLSC and NILC use different representation spaces, their absolute silhouette scores are not directly comparable. We therefore interpret silhouette changes within each method.

D.6 Evaluation Metrics

We assess clustering quality using four widely adopted metrics: Accuracy (ACC), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Fowlkes–Mallows Index (FMI). Higher scores correspond to better clustering performance for all four metrics.

Accuracy (ACC). ACC measures the best possible agreement between the predicted clusters and the ground-truth labels. Specifically, the predicted labels are optimally matched to the ground-truth labels using the Hungarian algorithm (Zhang et al., 2021a; 2024c), and ACC is then computed as the proportion of correctly assigned samples under this optimal permutation.

Normalized Mutual Information (NMI). NMI measures the amount of shared information between the predicted and ground-truth label distributions. The mutual information is normalized by the average entropy of the two distributions:

$$\mathrm{NMI}(y_{gt}, y_p) = \frac{\mathrm{MI}(y_{gt}, y_p)}{\frac{1}{2}\left[H(y_{gt}) + H(y_p)\right]}, \tag{22}$$

where y_{gt} and y_p denote the ground-truth and predicted labels, respectively, $\mathrm{MI}(\cdot)$ denotes mutual information, and $H(\cdot)$ denotes entropy.

Adjusted Rand Index (ARI). ARI evaluates the agreement between two assignments by examining all pairs of samples and correcting the resulting score for chance agreement. Let

$$S = \frac{\sum_i \binom{u_i}{2} \sum_j \binom{v_j}{2}}{\binom{n}{2}}. \tag{23}$$

The ARI is defined as

$$\text{ARI} = \frac{\sum_{i,j}\binom{n_{i,j}}{2} - S}{\frac{1}{2}\left[\sum_i \binom{u_i}{2} + \sum_j \binom{v_j}{2}\right] - S}, \tag{24}$$

where $n_{i,j}$ denotes the number of samples shared by the i-th and j-th clusters, and u_i and v_j are the marginal totals of the corresponding contingency table.

Fowlkes–Mallows Index (FMI). FMI summarizes clustering quality through the geometric mean of pairwise precision and recall:

$$\text{FMI} = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}}, \tag{25}$$

where TP, FP, and FN denote the numbers of true-positive, false-positive, and false-negative sample pairs, respectively.

E Additional Experimental Results

E.1 Additional Ablation Studies

Table 5: Ablation results at KCRs of 0.25 and 0.50. All scores are percentages averaged over five random seeds. Best and second-best results within each setting are **bold** and underlined, respectively.

KCR	Variant	MIntRec					MIntRec2.0					MELD-DA				
		ACC	ARI	NMI	FMI	Avg.	ACC	ARI	NMI	FMI	Avg.	ACC	ARI	NMI	FMI	Avg.
0.25	w/o $\mathcal{L}_{\text{pre}}$	32.76	14.58	39.96	19.88	26.80	21.29	10.70	30.89	14.79	19.42	23.81	6.97	12.01	20.72	15.88
	w/o $\mathcal{L}^{(r)}_{\text{refresh}}$	43.06	23.56	47.67	28.32	35.65	30.29	16.11	38.50	19.86	26.19	32.13	19.32	26.33	30.72	27.13
	w/o $\lambda^{(r)}_{\text{pce}}$ ramp-up	41.35	22.13	46.56	26.94	34.25	28.05	14.69	36.02	18.49	24.31	32.31	16.56	23.10	28.80	25.19
	w/o $\lambda_{\text{same}}, \lambda_{\text{different}}$	43.44	24.38	48.45	29.01	36.32	30.16	16.55	38.72	20.31	26.44	32.07	17.45	26.47	28.79	26.20
	w/o $\mathcal{L}^{(r)}_{\text{pck}}$	42.20	23.66	48.45	28.41	35.68	30.00	16.02	38.03	19.75	25.95	32.27	17.63	26.59	28.94	26.36
	PLSC	**43.73**	**24.41**	**48.46**	**29.11**	**36.43**	**30.54**	**16.84**	**38.99**	**20.58**	**26.74**	**35.56**	**20.42**	**26.97**	**31.86**	**28.70**
0.50	w/o $\mathcal{L}_{\text{pre}}$	35.60	18.10	42.65	23.13	29.87	26.17	14.61	35.58	18.48	23.71	27.19	10.85	17.19	23.69	19.73
	w/o $\mathcal{L}^{(r)}_{\text{refresh}}$	46.61	27.73	51.57	32.27	39.55	35.46	20.30	**42.16**	23.94	30.47	31.16	16.66	26.37	27.99	25.55
	w/o $\lambda^{(r)}_{\text{pce}}$ ramp-up	45.53	26.80	50.70	31.38	38.60	33.19	18.88	40.61	22.56	28.81	31.68	17.08	24.97	28.47	25.55
	w/o $\lambda_{\text{same}}, \lambda_{\text{different}}$	45.93	27.88	51.48	32.40	39.42	35.24	20.57	41.18	24.19	30.30	33.94	18.34	26.14	29.48	26.98
	w/o $\mathcal{L}^{(r)}_{\text{pck}}$	47.73	27.94	51.58	32.46	39.93	34.51	20.07	41.87	23.71	30.04	35.20	19.20	26.58	30.23	27.80
	PLSC	**47.78**	**28.90**	**52.01**	**33.39**	**40.52**	**35.72**	**20.77**	42.14	**24.39**	**30.76**	**35.50**	**19.53**	**26.69**	**31.01**	**28.18**

Table 5 presents ablation results at known class ratios of 0.25 and 0.50, complementing the analysis in Section 5.1. On MIntRec, MIntRec2.0, and MELD-DA, We evaluate the same five variants to assess their contributions when fewer intent categories provide labeled supervision.

w/o $\mathcal{L}_{\text{pre}}$. Removing semi-supervised multimodal pretraining produces the largest average decrease on every dataset at both ratios. At a KCR of 0.25, average scores decrease by 9.63%, 7.32%, and 12.82%, while at a KCR of 0.50 the decreases are 10.65%, 7.05%, and 8.45%. MELD-DA is particularly sensitive at lower known-class ratios, with both decreases exceeding the 4.19% drop at 0.75. These results emphasize the importance of establishing multimodal representations through cross-view consistency and known intent supervision before iterative clustering.

w/o $\mathcal{L}^{(r)}_{\text{refresh}}$. Omitting supervised contrastive learning on labeled samples before each clustering round reduces average scores by 0.78%, 0.55%, and 1.57% at a KCR of 0.25, and by 0.97%, 0.29%, and 2.63% at a KCR of 0.50. The largest decrease occurs on MELD-DA at both ratios. Although NMI on MIntRec2.0 increases slightly from 42.14% to 42.16% at a KCR of 0.50, its average score still declines. The overall pattern supports repeatedly using known intent labels to maintain semantic structure as representations evolve.

w/o $\lambda^{(r)}_{\text{pce}}$ ramp-up. Disabling the gradual increase in pseudo-label classification reduces average scores by 2.18%, 2.43%, and 3.51% at a KCR of 0.25, and by 1.92%, 1.95%, and 2.63% at a KCR of 0.50. For all three datasets, these decreases exceed those observed at 0.75. This pattern suggests that

gradually introducing pseudo-label supervision is particularly beneficial when fewer known intent categories are available, consistent with reducing the influence of unreliable early assignments.

w/o $\lambda_{\text{same}}, \lambda_{\text{different}}$. With both semantic penalties omitted and centroid optimization retained, average scores decrease by 0.11%, 0.30%, and 2.50% at a KCR of 0.25, and by 1.10%, 0.46%, and 1.20% at a KCR of 0.50. MELD-DA shows the largest decrease at both ratios, indicating greater sensitivity to removing the relational constraints in these settings. These results support incorporating MLLM judgments into centroid refinement to encourage semantically compatible representatives to share an assignment and incompatible ones to separate.

w/o $\mathcal{L}_{\text{pck}}^{(r)}$. Removing semantic centroid correction reduces average scores by 0.75%, 0.79%, and 2.34% at a KCR of 0.25, and by 0.59%, 0.72%, and 0.38% at a KCR of 0.50. The effect varies across datasets and KCRs, with the largest decrease on MELD-DA at a KCR of 0.25. Together with the results at a KCR of 0.75, these findings support initializing subsequent clustering rounds with corrected centroids, allowing local semantic relations to guide later assignments.

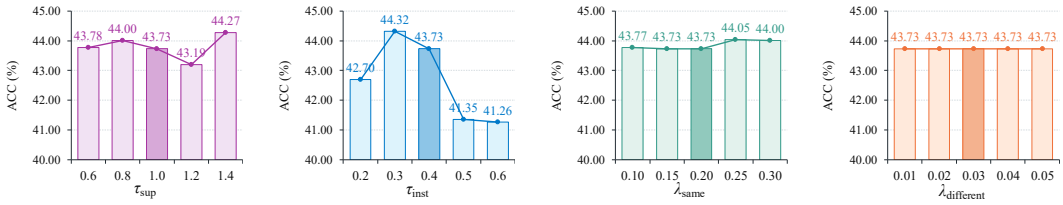

Figure 6: Sensitivity analysis of dataset-specific hyperparameters on MIntRec at a KCR of 0.25.

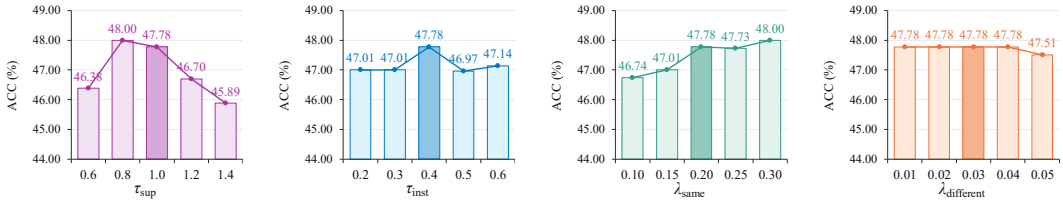

Figure 7: Sensitivity analysis of dataset-specific hyperparameters on MIntRec at a KCR of 0.5.

E.2 Additional Hyperparameter Sensitivity Analysis

Figures 6 and 7 extend the sensitivity analysis on MIntRec to KCRs of 0.25 and 0.50 under the same protocol as Section 5.3. At a KCR of 0.25, $\tau_{\text{sup}} = 1.4$ and $\tau_{\text{inst}} = 0.3$ achieve the highest ACC in their sweeps at 44.27% and 44.32%, compared with 43.73% under the default configuration. Increasing τ_{inst} to 0.6 reduces ACC to 41.26%, showing that the temperature governing cross-view consistency can substantially affect performance when fewer intent classes provide labeled supervision. Temperature controls the concentration of contrastive weighting among sample pairs, and its effect on shared representations can carry into the assignments used for subsequent pseudo-label learning. The semantic penalties vary less in this setting. ACC ranges from 43.73% to 44.05% as λ_{same} changes, while varying $\lambda_{\text{different}}$ leaves mean ACC unchanged at 43.73%. At a KCR of 0.50, the preferred supervised temperature shifts to $\tau_{\text{sup}} = 0.8$, which achieves 48.00%, only 0.22 percentage points above the adopted value of 1.0. Meanwhile, $\tau_{\text{inst}} = 0.4$ achieves the highest ACC of 47.78% within its sweep, and the variation across tested values narrows to 0.81 percentage points, compared with 3.06 percentage points at a KCR of 0.25. This difference shows that sensitivity to the instance contrastive temperature depends on the supervision setting. The SAME penalty displays the opposite pattern, with a wider ACC range at 0.50 than at 0.25. $\lambda_{\text{same}} = 0.10$ yields 46.74%, while 0.30 achieves 48.00%, and the default value of 0.20 attains 47.78%. A larger penalty is therefore beneficial in some settings, but its preferred strength varies with KCR, consistent with the need to balance semantic agreement between representatives against their distances to candidate centroids. The DIFFERENT penalty remains comparatively stable, yielding 47.78% from 0.01 to 0.04 and

decreasing slightly to 47.51% at 0.05. These broad plateaus suggest that aggregate performance is relatively insensitive to moderate changes in this penalty within the evaluated range. In PLSC, semantic relations enter centroid refinement as finite penalties alongside the feature distortion term, allowing their influence to be balanced with the multimodal representations. Although individual KCRs favor different parameter values, averaging ACC equally across all three KCRs places each adopted value first within its respective sweep, with $\lambda_{\text{different}} = 0.03$ tied with 0.04. This aggregate comparison supports the shared configuration as a favorable balance across supervision settings, while the individual results identify where temperature or constraint strength has a greater influence.

E.3 Additional Clustering Visualizations

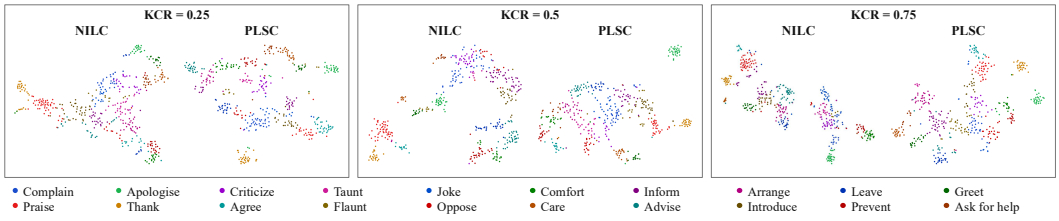

Figure 8: Visualization of representations on MIntRec across three KCRs, compared with NILC.

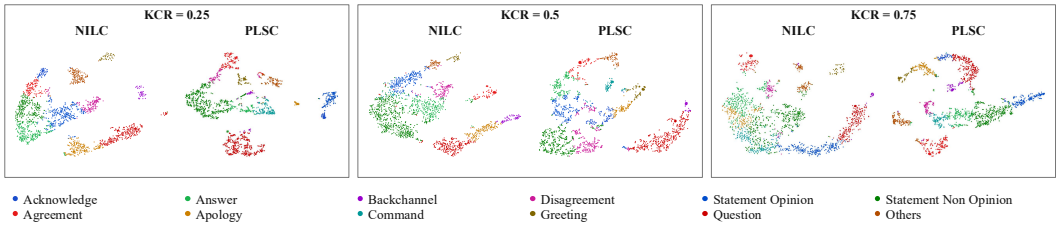

Figure 9: Visualization of representations on MELD-DA across three KCRs, compared with NILC.

Figures 8 and 9 extend the comparison with NILC to MIntRec and MELD-DA under three KCRs.

MIntRec. Figure 8 shows differences in both the concentration of individual intents and their overlap with neighboring categories. At a KCR of 0.25, *Thank* samples extend into a scattered tail in NILC, whereas PLSC gathers most of them into a compact group. At a KCR of 0.50, PLSC maintains distinct *Apologise* and *Thank* groups, although several other intents remain interleaved. A more revealing contrast appears at a KCR of 0.75, where *Complain*, *Criticize*, and *Taunt* occupy an overlapping region in NILC, while PLSC places most *Taunt* samples away from the other two. This distinction is particularly useful for separating related intents that can otherwise receive similar assignments. In iterative clustering, such mixed groups can produce pseudo labels that repeatedly reinforce the same ambiguity. Retaining known-intent supervision and gradually increasing the classification weight provide a way to maintain semantic distinctions while reducing the influence of early assignment errors. The remaining overlap between *Complain* and *Criticize* indicates that these benefits vary across categories, with the clearest improvements appearing in particular groups.

MELD-DA. Figure 9 reveals several cases where PLSC separates dialogue acts that are interleaved in NILC. At a KCR of 0.25, *Statement Opinion* overlaps with neighboring acts in NILC but occupies a largely separate region in PLSC, and *Apology* also forms a more concentrated group. At a KCR of 0.50, *Command* samples form a recognizable group in PLSC, while they are more dispersed among *Statement Opinion* and *Statement Non Opinion* in NILC. At a KCR of 0.75, the distinction between *Apology* and *Answer* becomes particularly clear, as their samples are extensively interspersed in NILC whereas PLSC places most *Apology* samples away from *Answer*. These examples highlight why proximity in the representation space may be insufficient to determine whether two samples express the same act. Comparing representatives through their original text and video gives PLSC additional evidence for discouraging incompatible assignments, which can influence later clustering

Table 6: Results on MIntRec2.0. PLSC$_{\widehat{K}}$ uses an estimated category count, while USNID and NILC use $K = 30$. Scores are percentages averaged over five seeds. Best and second-best results are **bold** and underlined, respectively.

Method	KCR = 0.25					KCR = 0.50					KCR = 0.75				
	ACC	ARI	NMI	FMI	Avg.	ACC	ARI	NMI	FMI	Avg.	ACC	ARI	NMI	FMI	Avg.
USNID	29.79	15.42	38.66	19.28	25.79	33.44	18.61	41.39	22.34	28.95	36.19	20.48	43.11	24.13	30.98
NILC	30.26	15.42	38.17	19.39	25.81	34.46	19.12	40.61	23.11	29.33	**38.32**	20.77	42.26	**25.15**	**31.63**
PLSC$_{\widehat{K}}$	**30.63**	**16.74**	**38.72**	**20.55**	**26.66**	**34.92**	**20.38**	**42.33**	**24.01**	**30.41**	35.76	**21.16**	**44.17**	24.81	31.48

Table 7: Detailed Runtime breakdown of PLSC on three benchmarks. All times are in seconds. Rep. denotes representative.

Dataset	Pre-training	SupCon refresh	Feature extraction	Clustering/ alignment	Density selection	Rep. selection	MLLM reasoning	PCK correction	Progressive training	Other	**Total**
MIntRec	268.76	31.61	83.26	1.06	8.41	0.19	625.72	0.49	158.67	13.93	**1192.10**
MIntRec2.0	1671.61	96.20	265.14	7.77	65.54	1.36	1553.75	0.37	862.28	34.76	**4558.78**
MELD-DA	943.51	142.10	329.97	3.79	107.53	0.90	1308.71	0.46	955.16	63.00	**3855.12**

through corrected centroids. Overlap among *Answer*, *Command*, and *Statement Non Opinion* nevertheless persists, and the visible improvements differ across ratios. The benefit is therefore most evident in resolving specific local ambiguities, including under sparse known-class supervision.

F Unknown Number of Intent Categories

Following prior work, our main experiments assume that the total number of intent categories K is known, ensuring consistent clustering granularity across methods. To assess performance when this information is unavailable, we estimate K from unlabeled representations using the strategy of USNID (Zhang et al., 2024c) and evaluate PLSC on MIntRec2.0, whose true category count is 30. This setting retains the labeled examples of known intents required for semi-supervised discovery but removes access to the total number of categories. For seeds 0–4, the estimated counts are $28, 30, 30, 27$, and 25 at a KCR of 0.25, $29, 29, 35, 34$, and 30 at 0.50, and $36, 32, 35, 34$, and 35 at 0.75, with corresponding means of 28.0, 31.4, and 34.4. These estimates include both underestimation and overestimation. The first setting predominantly underestimates the category count, the second produces estimates on both sides of the true value, and the third consistently overestimates it. We compare PLSC$_{\widehat{K}}$ with USNID and NILC using the true count $K = 30$, testing whether PLSC remains competitive when the reference methods retain this additional information.

As shown in Table 6, PLSC$_{\widehat{K}}$ achieves the highest scores across all four metrics at KCRs of 0.25 and 0.50. At 0.25, its average score reaches 26.66%, exceeding USNID and NILC by 0.87% and 0.85%, respectively. Compared with NILC, the improvements are particularly evident in ARI and FMI, which increase by 1.32% and 1.16%. Thus, moderate underestimation does not eliminate the advantage of PLSC in recovering relationships between samples from the same intent. At 0.50, the average score increases to 30.41%, outperforming USNID by 1.46% and NILC by 1.08%. PLSC$_{\widehat{K}}$ also achieves an NMI of 42.33%, compared with 41.39% for USNID and 40.61% for NILC. The gains across multiple metrics indicate that its advantage in these two settings is not confined to a single evaluation criterion. These observations are consistent with PLSC retaining known intent supervision throughout progressive learning while gradually increasing the contribution of pseudo-label classification, so representation learning does not depend entirely on assignments generated with an estimated category count. At a KCR of 0.75, the comparison becomes more mixed. PLSC$_{\widehat{K}}$ achieves the highest ARI and NMI, reaching 21.16% and 44.17%, respectively. Its average score of 31.48% remains 0.50% above USNID and only 0.15% below NILC, although its ACC is lower than both reference methods. The mean estimated count of 34.4 introduces more clusters than the true count of 30, making fragmentation of an intent across several predicted clusters a plausible explanation for the lower ACC. This interpretation is also consistent with the scope of semantic correction in PLSC, where relational judgments guide representative assignments and centroid refinement but the number of centroid slots remains fixed during correction. Consequently, semantic feedback can improve assignments within the estimated partition without directly recovering the true category count.

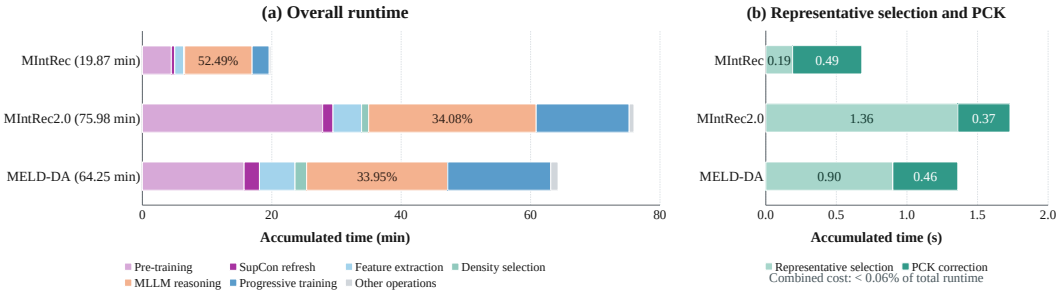

Figure 10: Time cost of PLSC on three benchmarks with a KCR of 0.5

The results therefore support the effectiveness of PLSC when K is estimated, while showing that persistent overestimation can narrow its advantage and affect different clustering metrics unevenly.

G Runtime and Resource Analysis

G.1 Time Efficiency

Figure 10 reports the runtime breakdown of PLSC on MIntRec, MIntRec2.0, and MELD-DA, with detailed timings in Table 7. Total times are 19.87, 75.98, and 64.25 minutes, respectively. MLLM reasoning accounts for 52.49%, 34.08%, and 33.95% of these totals, making it the largest individual component on MIntRec and MELD-DA. On MIntRec2.0, pretraining occupies a slightly larger share of 36.67%. Thus, although relative contributions vary across datasets, representation learning and MLLM reasoning consistently dominate the runtime.

Within representation learning, pretraining accounts for 22.55% to 36.67% of total runtime, while progressive training contributes 13.31% to 24.78%. These stages establish and update the multimodal representations using evolving clustering supervision. By comparison, the supervised contrastive refresh requires only 31.61, 96.20, and 142.10 seconds on the three datasets, corresponding to 2.65%, 2.11%, and 3.69% of total runtime. This modest cost shows that reinforcing known intent supervision before each clustering round does not become a major training overhead. Feature extraction contributes a further 5.82% to 8.56%, while clustering and alignment together require at most 7.77 seconds across the evaluated benchmarks. Once MLLM relations are available, incorporating them through numerical centroid correction is inexpensive. Representative selection takes 0.19, 1.36, and 0.90 seconds, while PCK correction requires 0.49, 0.37, and 0.46 seconds on MIntRec, MIntRec2.0, and MELD-DA, respectively. Their combined costs are therefore 0.68, 1.73, and 1.36 seconds, representing less than 0.06% of total runtime on every benchmark. This low cost is consistent with performing constrained refinement over selected representatives in a fixed feature space, without backpropagating through the multimodal encoder during correction. The corrected centroids initialize subsequent clustering rounds, allowing semantic relations to influence later assignments beyond the representatives used in correction. These results indicate that PLSC incorporates semantic guidance with an acceptable overall runtime under the evaluated settings. The relation cache enables repeated queries to reuse previously inferred judgments, reducing duplicate MLLM calls across clustering rounds. Once these judgments are available, their integration through numerical centroid correction incurs minimal additional computation. Together with the modest cost of supervised contrastive refresh, this allows PLSC to maintain known intent supervision and refine centroids throughout progressive learning. The corrected centroids subsequently guide new assignments, extending the influence of the inferred relations beyond the selected representatives. Overall, representative selection, cached relation reuse, and lightweight centroid refinement make iterative semantic correction practical within the complete learning pipeline.

G.2 Resource Consumption

Table 8 reports the MLLM query counts and token consumption of PLSC during semantic correction. To avoid repeated inference, we cache previously obtained relation judgments and retrieve

Table 8: MLLM resource consumption during semantic correction. New calls require MLLM inference, while cache hits reuse previously stored relation judgments.

Dataset	New calls	Cache hits	Input tokens	Output tokens
MIntRec	587	1,713	97,567	246,752
MIntRec2.0	2,141	1,334	7,712,190	815,829
MELD-DA	1,068	333	4,166,775	350,378

them whenever the same relation is queried again. New calls therefore represent queries requiring fresh MLLM inference, whereas cache hits reuse existing judgments without another MLLM invocation. This mechanism is particularly relevant to progressive learning, where representative relations may be revisited across clustering rounds.

PLSC records 587, 2,141, and 1,068 new calls on MIntRec, MIntRec2.0, and MELD-DA, respectively, alongside 1,713, 1,334, and 333 cache hits. Defining the cache hit rate as the number of cache hits divided by the sum of new calls and cache hits, the corresponding rates are 74.48%, 38.39%, and 23.77%. On MIntRec, nearly three quarters of the relation queries are resolved through the cache, indicating substantial reuse of previously inferred relations. Caching also avoids 1,334 and 333 fresh queries on the other two benchmarks. These results show that storing semantic judgments is useful throughout iterative correction, although the extent of reuse varies across datasets. Token consumption provides a complementary view of resource requirements. MIntRec uses 97,567 input tokens and 246,752 output tokens, compared with 7,712,190 and 815,829 on MIntRec2.0, and 4,166,775 and 350,378 on MELD-DA. The balance between input and output consumption differs considerably. Output tokens exceed input tokens on MIntRec, whereas input tokens account for most of the token usage on the other two datasets. Consequently, call counts alone do not fully characterize MLLM resource consumption, and the cache hit rates should be interpreted as query reuse rather than equivalent percentage reductions in token usage. By combining representative selection with the reuse of cached relations, PLSC directs fresh inference toward selected relations that have not already been evaluated, avoiding repeated MLLM processing of the same relation evidence. Overall, these results show that the resource cost of iterative semantic correction remains acceptable. Caching substantially reduces redundant MLLM calls, and fresh inference is directed only to relations not yet evaluated, avoiding repeated processing of the same semantic evidence. This makes iterative semantic correction practical within the complete learning pipeline.