

APPENDIX

In this appendix, we present additional related work, dataset descriptions, hyperparameter settings, experimental setup, comprehensive evaluation results, and extended experimental analyses. The detailed contents are organized as follows:

A	Related Works	15
A.1	Multimodal Representation Learning	15
A.2	Cross-Modal Alignment	15
B	Datasets	16
B.1	Motivation for Dataset Selection	16
B.2	Detailed Dataset Description	16
B.3	Hyper-parameter Settings	17
B.4	Feature Extraction	17
B.5	Evaluation Metric for IEMOCAP	18
B.6	Statistical Estimation of Modality-Common Features	18
C	Additional Experimental Analysis	18
C.1	Extended Analysis for CMU-MOSI and CMU-MOSEI	18
C.2	Extended Analysis for CH-SIMS	19
C.3	Extended Analysis for Ablation Studies	20
C.4	Extended Analysis for Modality Gap	23

A RELATED WORKS

A.1 MULTIMODAL REPRESENTATION LEARNING

Multimodal representation learning aims to integrate heterogeneous data from diverse modalities into a cohesive framework that captures complementary semantic information (Qian et al., 2025; Liang et al., 2024b; Bayoudh, 2024; Wang et al., 2025). Recent methods have achieved significantly performance improvements by leveraging representation-based and cross-modal interaction approaches. Specifically, Self-MM (Yu et al., 2021) applies self-supervised contrastive learning and masked modeling to enhance the mutual information across modalities, HGraph-CL (Lin et al., 2022) introduces hierarchical graph contrastive learning to model intricate interactions across modalities. However, the heterogeneity and complementary information inherent in multimodal representations are intrinsically entangled, making it challenging to fully harness their complementary strengths while preserving their unique characteristics. Inspired by this insight, MISA (Hazarika et al., 2020) separates multimodal representations into modality-invariant and modality-unique features with contrastive and reconstruction losses, DMD (Li et al., 2023) further introduces graph cross-modal knowledge distillation to explicitly model the correlations across modalities. However, existing methods are often constrained to modeling modalities from a global perspective, overlooking the token-level local semantic inconsistencies that arise in cross-modal interactions. Our proposed DecAlign enables fine-grained multimodal representation learning through hierarchical alignment, progressing from local to global and heterogeneity to homogeneity, ensuring precise cross-modal integration and semantic consistency.

A.2 CROSS-MODAL ALIGNMENT

The core challenge in multimodal tasks lies in the inherent heterogeneity across modalities (Zhu et al., 2024), characterized by structural, distributional, and semantic disparities, which restricts the

effective synergy of multimodal homogeneous features. To address this, existing solutions can be broadly categorized as follows: ❶ Shared Representation, which aims to learn a unified latent space for cross-modal semantic consistency. For example, CLIP-based methods (Radford et al., 2021; Gao et al., 2024) use contrastive learning to align image-text pairs in a shared embedding space, while Uni-Code (Xia et al., 2024) employs cross-modal information disentangling and exponential moving average to align semantically equivalent information in a shared latent space. ❷ Transformer-based methods that apply cross-attention to dynamically capture key information in cross-modal interactions (Tsai et al., 2019; Yang et al., 2022; Hu et al., 2024). ❸ Modality Translation, which establishes mappings between modalities through cross-modal generation or reconstruction (Liu et al., 2024b; Zeng et al., 2024; Tian et al., 2022). ❹ Cross-Modal Knowledge Distillation, which addresses inter-modal contribution imbalances and explores cross-modal correlations. For example, DMD (Li et al., 2023) employs graph distillation for dynamic knowledge transfer, and UMDf (Li et al., 2024a) uses unified self-distillation to learn robust representations from consistent multimodal distributions. Unlike methods that overemphasize homogeneous information, we tackle the issue of over-alignment diminishing modality-unique characteristics through representation decoupling and a hierarchical alignment mechanism, ensuring cross-modal semantic consistency while retaining unimodal characteristics.

B DATASETS

B.1 MOTIVATION FOR DATASET SELECTION

To comprehensively evaluate DecAlign’s effectiveness in multimodal sentiment analysis, we select four widely used benchmarking datasets. CMU-MOSI and CMU-MOSEI are well-established English multimodal sentiment datasets, enabling a direct and fair comparison with prior approaches. CH-SIMS extends our evaluation to a Chinese dataset, ensuring cross-linguistic generalization. Additionally, IEMOCAP (six-class) is included to assess the model’s performance in fine-grained emotion classification. By conducting experiments across datasets with different languages, sentiment labels, and scales, we provide a thorough assessment of DecAlign’s robustness and applicability.

Dataset	# Train	# Test	# Category	Modality		
				Audio	Visual	Text
CMU-MOSEI	16327	4659	2 & 7	✓	✓	✓
CMU-MOSI	1284	686	2 & 7	✓	✓	✓
CH-SIMS	1368	457	3	✓	✓	✓
IEMOCAP	5810	1623	6	✓	✓	✓

Table 3: Statistical information on four chosen datasets.

B.2 DETAILED DATASET DESCRIPTION

CMU-MOSI consists of 2,199 monologue movie review clips, each annotated with a sentiment score ranging from -3 (highly negative) to +3 (highly positive). It contains word-aligned multimodal signals, including textual, visual, and acoustic features. CMU-MOSI is commonly used for both sentiment classification and regression tasks, making it a crucial benchmark for evaluating multimodal models.

CMU-MOSEI extends CMU-MOSI by providing a significantly larger dataset with 22,856 opinion-based clips covering diverse topics, speakers, and recording conditions. Similar to CMU-MOSI, it includes multimodal data aligned at the word level and sentiment scores in the range of -3 to +3. Due to its large-scale and diverse nature, CMU-MOSEI is used to assess model generalization across various domains.

CH-SIMS consists of 38,280 Chinese utterances designed for multimodal sentiment analysis in Mandarin. Each sample includes textual, visual, and acoustic information, with sentiment labels ranging from -1 (negative) to +1 (positive). CH-SIMS enables research in cross-lingual sentiment analysis and serves as an important benchmark for multimodal sentiment models in Chinese contexts.

IEMOCAP (six-class). The IEMOCAP dataset comprises 10,039 dynamic utterances annotated with six emotion categories: angry, happy, sad, neutral, excited, and frustrated. Each sample contains textual, visual, and acoustic modalities. Due to its imbalanced class distribution, weighted accuracy (WAcc) and weighted average F1 score (WAF1) are commonly adopted to ensure fair performance evaluation in emotion recognition tasks.

B.3 HYPER-PARAMETER SETTINGS

Table 4 summarizes the dataset-specific hyperparameter configurations used in DECALIGN. We observe that while most parameters are kept consistent across datasets, several key components are tuned to accommodate dataset characteristics. For example, the dropout rates applied to different modalities (audio, visual, and text) vary slightly: MOSI and CH-SIMS adopt relatively higher dropout for text and output layers (0.4–0.6), reflecting their smaller dataset sizes and the need for stronger regularization, whereas MOSEI and IEMOCAP use more moderate dropout values (0.2) to balance regularization and information retention.

In terms of feature representation, the DST feature dimension and attention heads are adjusted according to dataset scale: MOSEI employs the largest dimension setting ([64, 8]), while IEMOCAP uses a smaller but deeper configuration ([48, 4]). Other architectural choices, such as Conv1D kernel size, transformer depth (4 layers), and batch size (32), remain uniform across all datasets to ensure comparability.

For optimization, the learning rate is tuned between 5×10^{-5} (MOSI and CH-SIMS) and 1×10^{-4} (MOSEI and IEMOCAP), with weight decay fixed at 0.005 and gradient clipping set between 0.5 and 0.6. The scheduler patience is consistently set to 5 across datasets. Regarding pretrained models, English datasets (MOSI, MOSEI, IEMOCAP) utilize `bert-base-uncased`, while the Chinese dataset CH-SIMS employs `bert-base-chinese`.

Table 4: Dataset-specific Hyperparameter Settings for DecAlign

Hyperparameter	MOSI	MOSEI	IEMOCAP	CH-SIMS
Attention Dropout (Audio)	0.3	0.2	0.2	0.3
Attention Dropout (Visual)	0.1	0.2	0.2	0.1
Attention Dropout (Text)	0.4	0.2	0.2	0.4
ReLU Dropout	0.1	0.2	0.2	0.1
Embedding Dropout	0.3	0.2	0.2	0.3
Residual Dropout	0.1	0.2	0.2	0.1
Output Dropout	0.6	0.2	0.2	0.6
Text Dropout	0.5	0.2	0.2	0.5
DST Feature Dim / Heads	[32, 8]	[64, 8]	[48, 4]	[32, 8]
Conv1D Kernel Size (L/A/V)	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5
Transformer Levels (nlevels)	4	4	4	4
Batch Size	32	32	32	32
Learning Rate	5e-5	1e-4	1e-4	5e-5
Weight Decay	0.005	0.005	0.005	0.005
Gradient Clipping	0.5	0.6	0.6	0.5
Scheduler Patience	5	5	5	5
Pretrained Model	<code>bert-base-uncased</code>	<code>bert-base-uncased</code>	<code>bert-base-uncased</code>	<code>bert-base-chinese</code>

B.4 FEATURE EXTRACTION

For IEMOCAP dataset, in line with previous studies [Lian et al. \(2023\)](#); [Fu et al. \(2024\)](#), we apply pre-trained DeBERTa [He et al. \(2020\)](#) to encode word sequences into 1024-dimensional texture embeddings for each utterance, while MA-Net [Zhao et al. \(2021\)](#) and wav2vec [Schneider et al. \(2019\)](#) are used to extract visual and acoustic features, respectively.

For CMU-MOSI, CMU-MOSEI, and CH-SIMS dataset, our multimodal feature extraction process is consistent with previous studies [Li et al. \(2023\)](#); [Wang et al. \(2023\)](#) by applying MMSA-FET Toolkit [Yu et al. \(2020; 2021\)](#) to extract features.

- **Text Modality:** For English datasets, we utilize the BERT-base-uncased model to extract 768-dimensional hidden states. For CH-SIMS, we apply a BERT-based-Chinese model. These pre-trained language models capture rich contextual semantics, improving sentiment representation.

- **Visual Modality:** We extract facial action features using the OpenFace toolkit’s Facet module, obtaining a 35-dimensional visual feature vector. These features capture facial expressions and microexpressions, which are essential for sentiment recognition.
- **Acoustic Modality:** We employ COVAREP to extract a 74-dimensional acoustic feature vector. These features include pitch, energy, and spectral properties, which are crucial for identifying speech-related emotional cues.

B.5 EVALUATION METRIC FOR IEMOCAP

To evaluate our proposed method on the IEMOCAP dataset, we adopt the following evaluation metrics. Denote C_0 as the number of emotion classes in the dataset, and Γ_j as the number of samples in class $j \in [1, C_0]$. Let Acc_j and $F1_j$ represent the classification accuracy and the F1 score of class j , respectively.

Weighted average accuracy (WAcc) is a weighted mean accuracy over different emotion classes, with weights proportional to the number of utterances in a particular emotion class. It is defined as:

$$WAA = \frac{\sum_{j=1}^{C_0} \Gamma_j \cdot Acc_j}{\sum_{j=1}^{C_0} \Gamma_j} \quad (14)$$

Similarly, weighted average F1 Score (WAF1) is a weighted mean F1 score over different emotion categories, using weights proportional to the number of utterances in each emotion class:

$$WAF1 = \frac{\sum_{j=1}^{C_0} \Gamma_j \cdot F1_j}{\sum_{j=1}^{C_0} \Gamma_j} \quad (15)$$

The IEMOCAP dataset consists of discrete emotion categories. To ensure a fair comparison with existing methods, we evaluate emotion recognition performance using both weighted average accuracy (WAcc) and weighted average F1-score (WAF1).

B.6 STATISTICAL ESTIMATION OF MODALITY-COMMON FEATURES

Mean: $\mu_{com}^{(m)} = \frac{1}{N} \sum_{n=1}^N \mathbf{f}_n^{(m)}$

Covariance: $\Sigma_{com}^{(m)} = \frac{1}{N} \sum_{n=1}^N (\mathbf{f}_n^{(m)} - \mu_{com}^{(m)})(\mathbf{f}_n^{(m)} - \mu_{com}^{(m)})^\top$

Skewness: $\Gamma_{com}^{(m)} = \frac{1}{N} \sum_{n=1}^N \left(\frac{\mathbf{f}_n^{(m)} - \mu_{com}^{(m)}}{\sqrt{\text{diag}(\Sigma_{com}^{(m)}) + \epsilon}} \right)^3$

C ADDITIONAL EXPERIMENTAL ANALYSIS

C.1 EXTENDED ANALYSIS FOR CMU-MOSI AND CMU-MOSEI

Tables 5 and 6 present the extended comparison results on the CMU-MOSI and CMU-MOSEI datasets. Several insightful trends can be observed.

Early Fusion and Transformer-based Methods. Early multimodal fusion approaches such as MFM (Tsai et al., 2018) and MulT (Tsai et al., 2019) achieve competitive baseline results, demonstrating the effectiveness of early cross-modal attention mechanisms. However, these models exhibit limitations in capturing fine-grained interactions, particularly when modality-specific features interfere with global semantics. For example, MulT improves over MFM by explicitly modeling directional cross-attention, yet still suffers from modality imbalance and over-reliance on dominant modalities.

Feature Decoupling and Disentanglement. Subsequent models such as MISA (Hazarika et al., 2020), CENet (Wang et al., 2022), and Self-MM (Yu et al., 2021) introduce feature disentanglement or modality-invariant modeling to alleviate semantic interference. By separating modality-common and

modality-specific features, these approaches achieve steady gains in both correlation and classification metrics. For instance, Self-MM attains a relatively high Corr (0.764) and F1 score (83.04) on MOSI, while FDMER (Yang et al., 2022) further enhances disentangled representation learning through factorized modeling, pushing the F1 score to 83.22.

Pre-trained Models and Advanced Alignment. Recent advances such as AOBERT (Kim & Park, 2023), DMD (Li et al., 2023), and CGGM (Guo et al., 2025) leverage pre-trained language models and sophisticated alignment strategies. AOBERT benefits from the representational power of BERT for textual features, while CGGM integrates classifier-guided generative modeling to better capture distributional patterns. DMD, in particular, achieves strong results across both regression and classification tasks, with MOSI (MAE = 0.744, Acc-2 = 83.24, F1 = 83.55) and MOSEI (MAE = 0.561, Acc-2 = 84.17, F1 = 83.88), showing the advantage of graph-based knowledge distillation for balancing modality-specific and modality-shared information.

Our DecAlign. Most importantly, our proposed DecAlign consistently outperforms all baselines by a notable margin. On MOSI, DecAlign achieves the lowest MAE (0.735), the highest Corr (0.811), and significant improvements in both Acc-2 (85.75) and Acc-7 (45.07), as well as F1 score (85.82). Similarly, on MOSEI, DecAlign sets new benchmarks with MAE = 0.543, Corr = 0.768, Acc-2 = 86.48, Acc-7 = 55.02, and F1 = 86.07. These improvements of around 2–3 points in Acc-2 and F1 score compared to the strongest baselines (e.g., DMD and CGGM) highlight three key strengths: ❶ the ability of prototype-guided optimal transport to capture fine-grained heterogeneous interactions, ❷ the effectiveness of hierarchical alignment in balancing modality-common and modality-unique representations, and ❸ improved robustness against modality imbalance, enabling DecAlign to model subtle multimodal signals more faithfully.

Models	MAE (↓)	Corr (↑)	Acc-2 (↑)	Acc-7 (↑)	F1 Score (↑)
MFM Tsai et al. (2018)	0.951	0.662	78.18	36.21	78.10
MuT Tsai et al. (2019)	0.846	0.725	81.70	40.05	81.66
PMR Fan et al. (2023)	0.895	0.689	79.88	40.60	79.83
CubeMLP Sun et al. (2022)	0.838	0.695	81.85	41.03	81.74
MUTA-Net Tang et al. (2023)	0.767	0.736	82.12	40.88	82.07
MISA Hazarika et al. (2020)	0.788	0.744	82.07	41.27	82.43
CENet Wang et al. (2022)	0.745	0.749	82.40	41.32	82.56
Self-MM Yu et al. (2021)	0.765	0.764	82.88	42.03	83.04
FDMER Yang et al. (2022)	0.760	0.777	83.01	42.88	83.22
AOBERT Kim & Park (2023)	0.780	0.773	83.03	43.21	83.02
DMD Li et al. (2023)	<u>0.744</u>	0.788	<u>83.24</u>	<u>43.88</u>	<u>83.55</u>
ReconBoost Hua et al. (2024)	0.793	0.769	82.59	42.70	82.72
CGGM Guo et al. (2025)	0.787	<u>0.792</u>	82.73	43.47	82.89
DecAlign (Ours)	0.735	0.811	85.75	45.07	85.82

Table 5: Performance Comparison on CMU-MOSI dataset. ↑ and ↓ indicate that higher or lower value is better. Best results are highlighted in **bold**, and suboptimal results are underlined. All reported results are averaged over **five** runs on the test set.

C.2 EXTENDED ANALYSIS FOR CH-SIMS

Table 7 reports the results on the CH-SIMS dataset (Yu et al., 2020), which is particularly challenging due to its Chinese language modality and the greater diversity of sentiment expressions. Several key observations can be made.

Early Fusion and Transformer-based Models. Traditional fusion models such as MFM (Tsai et al., 2018) and MuT (Tsai et al., 2019) provide reasonable baselines, showing that cross-attention mechanisms are effective in modeling interactions across modalities. However, their performance is limited by over-reliance on direct fusion, which often struggles to capture the nuanced sentiment variations inherent in Chinese multimodal data.

Disentanglement and Modality-specific Modeling. Models such as MISA (Hazarika et al., 2020), CENet (Wang et al., 2022), and FDMER (Yang et al., 2022) achieve steady improvements by explicitly modeling modality-invariant and modality-specific features. This disentanglement reduces semantic

Models	MAE ($\downarrow$)	Corr ($\uparrow$)	Acc-2 ($\uparrow$)	Acc-7 ($\uparrow$)	F1 Score ($\uparrow$)
MFM Tsai et al. (2018)	0.681	0.555	78.93	45.93	76.45
MuT Tsai et al. (2019)	0.673	0.677	80.85	48.37	80.86
PMR Fan et al. (2023)	0.645	0.689	81.57	48.88	81.56
CubeMLP Sun et al. (2022)	0.601	0.701	81.36	49.07	81.75
MUTA-Net Tang et al. (2023)	0.617	0.717	81.76	49.88	82.01
MISA Hazarika et al. (2020)	0.594	0.724	82.03	51.43	82.13
CENet Wang et al. (2022)	0.588	0.738	82.13	52.31	82.35
Self-MM Yu et al. (2021)	0.576	0.732	82.43	52.68	82.47
FDMER Yang et al. (2022)	0.571	0.743	83.88	53.21	83.35
AOBERT Kim & Park (2023)	0.588	0.738	83.90	52.47	83.14
DMD Li et al. (2023)	<u>0.561</u>	0.758	<u>84.17</u>	<u>54.18</u>	83.88
ReconBoost Hua et al. (2024)	0.599	0.733	82.98	52.68	83.14
CGGM Guo et al. (2025)	0.584	<u>0.760</u>	83.72	52.88	<u>83.94</u>
DecAlign (Ours)	0.543	0.768	86.48	55.02	86.07

Table 6: Performance Comparison on CMU-MOSEI dataset. $\uparrow$ and $\downarrow$ indicate that higher or lower value is better. Best results are highlighted in **bold**, and suboptimal results are underlined. All reported results are averaged over **five** runs on the test set.

interference and enables more precise sentiment prediction. For example, FDMER demonstrates robust performance with reduced MAE and improved F1, confirming the benefits of factorized disentanglement for sentiment modeling.

Advanced Alignment and Classifier-guided Fusion. Recent methods such as DMD (Li et al., 2023), MCIS (Zhou et al., 2025), and CGGM (Guo et al., 2025) deliver stronger baselines by leveraging more sophisticated alignment strategies. DMD and MCIS achieve MAE around 0.421–0.429 with F1 scores close to 80, showing balanced regression and classification performance. Notably, CGGM achieves the best results among existing baselines (MAE = 0.417, F1 = 80.12, Acc-3 = 80.17, Corr = 0.638), highlighting the effectiveness of classifier-guided generative modeling in capturing sentiment-related cues in Chinese multimodal contexts.

Our DecAlign. Despite these advances, DecAlign consistently achieves superior performance across all metrics, with MAE = 0.403, F1 = 81.85, Acc-3 = 88.24, and Corr = 0.657. These gains are substantial: in particular, the +8 point improvement in Acc-3 compared to the strongest baseline (CGGM) underscores DecAlign’s ability to achieve fine-grained categorical sentiment classification. The consistent reduction in MAE and higher correlation also indicate that DecAlign better captures continuous sentiment intensity.

Key Insights. The results validate three important properties of DecAlign: ❶ *Cross-lingual robustness*: by successfully handling Chinese sentiment data, DecAlign demonstrates strong generalization beyond English benchmarks. ❷ *Hierarchical alignment effectiveness*: prototype-guided optimal transport and latent semantic consistency work synergistically to reduce modality interference and enhance categorical prediction. ❸ *Balanced regression and classification*: DecAlign achieves state-of-the-art results in both continuous and discrete tasks, reflecting its ability to model both sentiment strength and polarity.

C.3 EXTENDED ANALYSIS FOR ABLATION STUDIES

We provide an in-depth analysis of the ablation results in Table 2 and the visualizations in Figure 4. We examine both the impact of *key components*—Multimodal Feature Decoupling (MFD), Heterogeneity Alignment (Hete), and Homogeneity Alignment (Homo)—and the contribution of *specific strategies* within the alignment modules, namely Prototype-guided Optimal Transport (Proto-OT), Contrastive Training (CT), Semantic Consistency (Sem), and Maximum Mean Discrepancy regularization (MMD).

A. Key components: MFD, Hete, Homo.

❶ *Decoupling is a prerequisite, alignment is the catalyst.* When only MFD is retained (i.e., both Hete and Homo are removed), the model degrades substantially (MOSI: MAE 0.784, F1 81.92;

Models	MAE ($\downarrow$)	F1 Score ($\uparrow$)	Acc-3 ($\uparrow$)	Corr ($\uparrow$)
MFM Tsai et al. (2018)	0.471	75.28	75.32	0.516
MuT Tsai et al. (2019)	0.455	76.96	77.02	0.544
PMR Fan et al. (2023)	0.445	76.55	76.63	0.523
CubeMLP Sun et al. (2022)	0.459	77.85	77.94	0.562
MUTA-Net Tang et al. (2023)	0.443	77.21	77.44	0.573
MISA Hazarika et al. (2020)	0.437	78.43	78.56	0.581
CENet Wang et al. (2022)	0.454	78.03	78.15	0.589
Self-MM Yu et al. (2021)	0.432	77.97	78.03	0.593
FDMER Yang et al. (2022)	0.424	78.74	78.82	0.599
AOBERT Kim & Park (2023)	0.430	78.55	78.65	0.578
DMD Li et al. (2023)	0.421	79.88	78.98	0.612
MCIS Yang et al. (2024)	0.429	79.58	79.64	0.629
CGGM Guo et al. (2025)	<u>0.417</u>	<u>80.12</u>	<u>80.17</u>	<u>0.638</u>
DecAlign (Ours)	0.403	81.85	88.24	0.657

Table 7: Performance comparison on the CH-SIMS dataset. $\uparrow$ indicates that higher values are better. Best results are highlighted in **bold**, and runner-up results are underlined.

MOSEI: MAE 0.632, F1 82.22). This indicates that decoupling alone, while preventing raw feature entanglement, cannot suppress modality-unique interference nor enforce cross-modal semantic coherence. Enabling both alignments in the full model reduces error and boosts classification notably (MOSI: MAE **0.735**, F1 **85.82**; MOSEI: MAE **0.543**, F1 **86.07**), yielding absolute gains over the MFD-only setting of $\Delta\text{MAE}=0.049/0.089$ and $\Delta\text{F1}=3.90/3.85$ on MOSI/MOSEI, respectively. This establishes that *decoupling prepares the space, while alignment shapes it*.

② *Heterogeneity alignment drives regression error down; homogeneity alignment stabilizes classification.* With **Hete**-only (no Homo), the model achieves lower MAE and higher F1 than Homo-only: MOSI (MAE 0.747, F1 84.46) vs Homo-only (MAE 0.754, F1 84.03); MOSEI (MAE 0.562, F1 84.74) vs Homo-only (MAE 0.588, F1 84.37). This shows that addressing *distributional mismatch* via Hete (Proto-OT + transformer refinement) has a more direct impact on minimizing continuous error (MAE), while Homo alignment (Sem + MMD) contributes more to *global semantic smoothing and decision stability*. In short, Hete is the primary lever for regression fidelity; Homo ensures calibrated and consistent boundaries.

③ *Hierarchical complementarity is essential, especially for subtle sentiments.* Figure 4(a)–(d) reveals that removing either Hete or Homo produces uneven drops across categories, with pronounced degradation at $\{-1, 0, +1\}$ sentiment levels. This pattern suggests that the two alignments address complementary failure modes: Hete mitigates local structure mismatches (reducing noisy shifts around decision margins), while Homo enforces global semantic agreement (preventing over-fragmented boundaries). Their combination recovers both *fine-grained discrimination* and *global robustness*.

④ *From modality gap to semantic co-location.* The t-SNE visualizations in Figure 4(e)–(h) corroborate the above: without alignment, paired language–vision features are distant with erratic pairwise directions; with Homo-only, clusters get closer but remain fragmented; with Hete-only, pairwise distances shrink further but residual anisotropy persists. Only the full model yields *tight, co-located* clusters, reflecting both local and global alignment. This explains the joint improvements in MAE and F1.

B. Strategy-level analysis: Proto-OT, CT, Sem, MMD.

① *Prototype-guided Optimal Transport (Proto-OT) is the backbone for error reduction.* Removing Proto-OT leads to marked MAE increases (MOSI: 0.748; MOSEI: 0.624 vs full 0.735/0.543), and consistent F1 drops (MOSI: 84.17, MOSEI: 85.03). This shows that *distribution-aware, prototype-level* alignment is indispensable for resolving cross-modal heterogeneity, especially in regression where small misalignments accumulate into larger intensity errors.

② *Contrastive Training (CT) is crucial for discriminability and margin preservation.* Without CT, inter-class separation weakens (MOSI F1: 84.36; MOSEI F1: 85.21), and MAE also deteriorates (MOSI: 0.743, MOSEI: 0.619). CT establishes *class-aware* anchors in the aligned space, preventing representation collapse and ensuring sharper decision boundaries—an effect that directly benefits Acc/F1 while indirectly stabilizing MAE by discouraging ambiguous embeddings near thresholds.

③ *Semantic Consistency (Sem) aligns moments beyond the mean—vital for stable fusion.* Replacing Sem with only MMD (i.e., removing the explicit latent-moment matching) yields broader, more variable clusters and a noticeable performance drop (MOSI F1 84.73 vs full 85.82; MOSEI F1 85.33 vs full 86.07). Sem’s explicit constraints on mean, covariance, and higher-order structure improve *global shape agreement* across modalities, thereby regularizing fusion and mitigating systematic biases.

④ *MMD supplies non-parametric distributional regularization that smooths the latent geometry.* Omitting MMD also harms performance (MOSI: MAE 0.741, F1 84.61; MOSEI: MAE 0.564, F1 85.26). While its marginal effect may appear smaller than Proto-OT or CT, MMD complements Sem by *non-parametrically* matching distributions in RKHS, capturing high-order statistics and preventing overfitting to specific batch-level alignments.

⑤ *Orthogonal roles, additive gains.* Comparing the four single-removal cases with the full model (MOSI: 0.735/85.82; MOSEI: 0.543/86.07), removing Proto-OT or CT most strongly hurts MAE and F1 (structure and margin), whereas removing Sem or MMD still impairs performance (global coherence and smoothness) but less severely. This ordering suggests that *structural alignment (Proto-OT)* and *discriminative supervision (CT)* form the core, while *semantic moment matching (Sem)* and *non-parametric regularization (MMD)* supply the necessary global consistency to fully realize the gains.

C. Cross-metric and cross-dataset insights.

① *MAE vs F1: different facets of the same alignment.* Hete (via Proto-OT) predominantly reduces MAE, reflecting improved geometric co-registration of modality-unique structures; Homo (Sem+MMD) mainly consolidates F1 by smoothing the cross-modal manifold and calibrating decision surfaces. Their synergy explains the concurrent improvements in both regression and classification.

② *Dataset scale and heterogeneity matter.* On the larger and more diverse MOSEI, removing CT or Proto-OT incurs greater MAE penalties (+0.076/+0.081) than on MOSI (+0.008/+0.013), indicating that *class-aware structure* and *prototype-level transport* are especially critical when the data distribution is broader and more multimodal.

③ *Category sensitivity and boundary sharpening.* Figure 4(a)–(d) shows that neutral and near-neutral bins suffer most when either alignment is removed. This suggests that hierarchical alignment is particularly effective at *sharpening ambiguous boundaries*, where cross-modal cues are subtle and easily swamped by dominant modality noise.

④ *From local structure to global semantics.* The combined evidence from Table 2 and Figure 4 indicates a two-stage mechanism: Hete reduces local structural discrepancies (prototype geometry, density mismatch), while Homo imposes global semantic consistency (moment alignment, RKHS discrepancy). The full model closes the *local-to-global* loop, yielding robust improvements across all metrics.

D. Takeaways for designing multimodal aligners.

① *Always decouple before you align.* MFD isolates modality-unique and -common factors, ensuring that subsequent alignment targets the right subspaces instead of wrestling with entangled representations.

② *Prioritize structure-aware alignment.* Prototype-level transport should be a first-class component: it is consistently the strongest driver of MAE reduction and a key stabilizer of downstream classification.

③ *Do not trade off discriminability for alignment.* Contrastive supervision is necessary to maintain class margins during alignment; it prevents over-alignment that collapses inter-class structure.

④ *Match distributions twice: parametrically and non-parametrically.* Moment-based Sem and kernel-based MMD play distinct roles and are most effective in tandem, aligning both *shape* and *support* of latent distributions.

Overall, the ablation evidence supports DecAlign’s hierarchical design: structural alignment of heterogeneous factors (Proto-OT + transformer refinement) combined with semantic and distributional alignment of homogeneous factors (Sem + MMD) is *jointly necessary* to achieve the consistent, cross-metric improvements observed across benchmarks.

C.4 EXTENDED ANALYSIS FOR MODALITY GAP

Figure 4 (e)–(h) visualizes the modality gap between language and vision features under different ablation settings. Each subplot presents paired features from the two modalities after projection into a 2D space using t-SNE, where lines connect corresponding language–vision pairs of the same input instance. Several key insights can be drawn.

Lack of Alignment. In subfigure (e), where both heterogeneity and homogeneity alignment modules are removed, the projected features from language and vision are widely separated with irregular pairwise connections. The high dispersion and inconsistent alignment directions indicate a severe modality gap, driven by the absence of constraints on either modality-unique discrepancies or cross-modal commonality. This validates that simple multimodal feature projection is insufficient for achieving semantic consistency.

Effect of Homogeneity-only Alignment. Subfigure (f), which removes the heterogeneity alignment but preserves homogeneity alignment, shows partial improvements: paired features are closer, and cluster overlap increases. However, disjoint sub-clusters remain visible, suggesting that enforcing semantic consistency via latent distribution matching reduces global misalignment but fails to resolve modality-specific variations. This highlights that semantic alignment alone cannot fully mitigate distributional heterogeneity.

Effect of Heterogeneity-only Alignment. In subfigure (g), where homogeneity alignment is removed but heterogeneity alignment is retained, the features are more concentrated and inter-modal distances shrink further. Prototype-guided optimal transport effectively aligns modality-unique structures by reducing distributional mismatch. Nevertheless, residual vertical dispersion across clusters reveals that without semantic alignment, global consistency is not guaranteed, leaving subtle but systematic modality biases uncorrected.

Full DecAlign with Hierarchical Alignment. Finally, subfigure (h) illustrates the full DecAlign framework with both alignment strategies enabled. Here, paired features are tightly clustered and nearly co-located, with minimal alignment distances and consistent cluster structures across modalities. This demonstrates the complementary effects of heterogeneity and homogeneity alignment: the former resolves modality-specific discrepancies at the distributional level, while the latter enforces semantic consistency in the latent space. Their combination closes the modality gap both locally and globally, leading to highly consistent cross-modal representations.

Key Insights. These visual analyses validate that *hierarchical alignment is crucial for robust multimodal integration*. Removing either module leads to partial but incomplete alignment, while the joint application of both substantially minimizes the modality gap. This explains why DecAlign achieves significant gains across benchmarks: it harmonizes modality-unique and modality-common representations simultaneously, ensuring that cross-modal signals are both semantically consistent and structurally coherent.