Evolutionary Multimodal Reasoning via Hierarchical Semantic Representation for Intent Recognition

Supplementary Material

1. Related Work for MLLMs

Multimodal Large Language Models (MLLMs) extend traditional Large Language Models (LLMs) by incorporating multimodal reasoning capabilities, facilitating integrated understanding across text, vision, and other modalities. Early approaches can be broadly categorized into two paradigms, including tool-integrated frameworks and end-to-end architectures. Tool-integrated methods, such as MM-REACT [20] and HuggingGPT [15], employ LLMs as central coordinators that invoke external vision models based on specific task requirements, demonstrating the modularity and extensibility of this approach. In contrast, end-to-end architectures aim to learn unified multimodal representations within a single model. Notable examples include Flamingo [2], which integrates frozen LLMs and vision encoders through gated cross-attention, and BLIP-2 [9], which introduces a Q-Former to effectively align visual and textual features. Recent advances in MLLMs emphasize enhanced reasoning, scalability, and generalization across diverse modalities. For example, Qwen2-VL [18] and LLaVA-NEXT [23] adopt dynamic resolution and linear scaling strategies for efficient video understanding. LLaMA 4 [1] introduces adaptive multimodal fusion mechanisms to support modality-aware representation learning, while NVLM 1.0 [4] applies contrastive pretraining to enhance cross-modal alignment. In parallel, Liquid [19] discretizes visual inputs into tokenized embeddings, enabling unified visual-text generation, and HyperLLaVA [22] employs adaptive tuning of both the projector and LLM parameters, alongside dynamic experts for visual and language processing, to enhance fine-grained alignment and optimize multimodal interactions. Furthermore, Chain-of-Thought (CoT) reasoning has emerged as a key enhancement for MLLMs, with methods like Visual CoT [14] offering superior problem decomposition and cognitive simulation, further strengthening the multimodal reasoning capabilities.

2. Supplementary Method Details

2.1. Details of Prompt Design

This section presents a representative example of the CoT prompt used in HIER on the MIntRec2.0 dataset and outlines the core principles behind its design. To explicitly demonstrate the hierarchical multimodal reasoning process, we construct a structured prompt that aligns with the semantic stages of HIER. This prompt guides the model through a progressive reasoning flow, simulating human-like cognitive steps for intent understanding. By organizing intent recognition into distinct stages of contextual grounding, concept analysis, and relation reasoning, the prompt enables more interpretable, accurate, and semantically coherent predictions.

Prompt Example from MIntRec2.0

Instruction: You are tasked with multimodal intent recognition, which involves analyzing both the visual content of a video and its corresponding text captions to determine the speaker's intent. Your goal is to select the most appropriate intent label from the provided list of thirty options: ['Acknowledge', 'Advise', 'Agree', 'Apologise', 'Arrange', 'Ask for help', 'Asking for opinions', 'Care', 'Comfort','Complain', 'Confirm', 'Criticize', 'Doubt', 'Emphasize', 'Explain', 'Flaunt','Greet','Inform', 'Introduce', 'Invite', 'Joke', 'Leave', 'Oppose', 'Plan', 'Praise', 'Prevent', 'Refuse', 'Taunt', 'Thank', 'Warn'].
CoT: The video is `<video>` and the text is `<text_start>` hey. don't distract jonah. `<text_end>`. Follow these steps: (1) Identify the main context by analyzing the video and summarizing the text captions; (2) Analyze key concepts `<concept>` relevant to the context and judge their usefulness; (3) Determine the intent label by considering the relations `<relation>` between key concepts and judge their usefulness. Use chain-of-thought reasoning to justify your decision, ensuring clarity and accuracy in your analysis. The output should be a single intent label.

Specifically, the prompt consists of detailed reasoning instructions and a carefully designed CoT. The instructions focus on defining the task and clarifying its objective, clearly introducing multimodal intent classification and providing a list of target intent labels. This helps the model understand the scope of the task and establishes a consistent framework for reasoning. In addition, the CoT design emulates the human reasoning process, guiding the MLLM to engage with and reason over the hierarchical semantic representations generated by the algorithm. Each input feature, including video, text tokens, concept tokens, and relation tokens, is clearly marked with distinct special symbols in accordance with the configuration used by Qwen2-VL. This labeling facilitates accurate substitution of representations and supports effective self-refinement during the subsequent reasoning process. In CoT-1, the model performs contextual grounding by integrating visual content with textual information. CoT-2 focuses on analyzing concepts extracted through multimodal concept clustering. CoT-3 involves structured reasoning over relation-

Table 1. Training and inference cost of HIER and its ablations on the MIntRec2.0 dataset, including Training Parameters (M), GPU Memory Usage (GB), Training Time (h/epoch), Inference Latency (s/sample), and Throughput (tokens/s).

Methods	Training Parameters	GPU Memory Usage	Training Time	Inference Latency	Throughput
HIER	20.37	38.17	1.32	1.04	432.69
w/o Concept	20.36	37.65	1.29	0.96	468.75
w/o Relation	20.19	36.44	1.25	0.87	517.24
w/o Self-evolution	20.34	38.03	1.27	0.94	478.72
Qwen2-VL	20.19	36.21	1.16	0.81	555.56

Table 2. Main parameters of HIER on the MIntRec, MIntRec2.0, and MELD-DA datasets.

Datasets	Learning Rate	Batch Size	Concept Tokens	Relation Tokens	Epochs	Accumulation	Calculation Precision
MIntRec	5e-5	4	50	25	5	8	bf16
MIntRec2.0	5e-5	4	40	20	5	8	bf16
MELD-DA	1e-4	4	40	20	5	8	bf16

ships between concepts, based on outputs from relation extraction modules. This staged reasoning approach fosters hierarchical semantic abstraction, contributing to more interpretable and semantically grounded predictions. To further enhance the quality of reasoning, reflective prompting is incorporated into both CoT-2 and CoT-3. Inspired by recent advances in self-evolution prompting, this prompt includes self-assessment *judge their usefulness*, encouraging the model to critically evaluate the relevance and informativeness of the extracted semantics. This promotes more deliberate and effective reasoning throughout the process.

2.2. Details of Spherical K-Means++

In this section, we provide a more detailed description of the Spherical K-Means++ algorithm employed for semantic concept clustering. Unlike standard K-Means++, which relies on Euclidean distance and is sensitive to vector magnitudes, Spherical K-Means++ measures similarity using cosine distance, effectively capturing semantic closeness in high-dimensional embedding spaces [16]. This makes it suitable for clustering multimodal tokens, where the direction of feature vectors often conveys more meaningful information than their magnitude.

Given the input token set Z, the objective is to partition these tokens into k semantic clusters $\{C_1, C_2, \ldots, C_k\}$, where each cluster is associated with a centroid vector $c_i \in \mathbb{R}^d$. The clustering process begins with centroid initialization, where the first centroid $c_1^{(0)}$ is randomly selected from Z. For each subsequent centroid $c_m^{(0)}$, the probability $p_i^{(0)}$ for token z_i is computed based on its minimum cosine distance to previous centroids:

$$p_i^{(0)} = \frac{D(z_i)}{\sum_{z \in Z} D(z)}, \; D(z) = \min_{1 \le j < s} (1 - \cos(z, c_j))^2, \quad (1)$$

where $\cos(\cdot, \cdot)$ denotes the calculation of cosine similarity, and s is the current number of selected centroids. Then, the new centroid $c_m^{(0)}$ is initialized as the weighted sum of all tokens $\sum_{i=1}^{n} p_i^{(0)} z_i$. At each iteration, token-to-centroid assignments are computed using a softmax over cosine similarities, allowing each token to contribute fractionally to multiple centroids. For example, during the the u-th iteration, the assignment probability $p_{i,m}^{(u)}$ is formalized as

$$p_{i,m}^{(u)} = \frac{\exp(\cos(z_i, c_m^{(u-1)}))}{\sum_{j=1}^{k} \exp(\cos(z_i, c_j^{(u-1)}))}, \quad (2)$$

where z_i is the i-th token and $c_m^{(u-1)}$ is the m-th centroid generated by the previous iteration. This assignment probability is then leveraged to update the centroid via $c_m^{(u)} = \sum_{i=1}^{n} p_{i,m}^{(u)} z_i$. By employing cosine-based similarity and soft assignments, Spherical K-Means++ enables differentiable clustering that aligns closely with the geometric properties of token embeddings, making it well-suited for multimodal semantic modeling.

3. Extended Experimental Settings

3.1. Training and Inference Cost

To further evaluate the practical viability of HIER for real-world deployment, we assess both its training and inference costs under consistent and realistic experimental conditions. Our objective is to determine whether HIER can deliver enhanced reasoning capabilities while maintaining computational efficiency comparable to existing MLLMs. We report five key indicators including the number of trainable parameters, peak GPU memory usage during training, training time per epoch, inference latency measured by the time required to process a single sample, and inference throughput measured in tokens per second. To ensure a fair comparison in computational efficiency, all experiments are conducted

on a single NVIDIA A100-PCIE GPU with 40 GB of memory. The results are presented in Table 1.

Compared to Qwen2-VL, HIER incurs only marginal computational overhead. It comprises 20.37 million trainable parameters, only 0.18 million more than Qwen2-VL, indicating a negligible increase of less than 1% in model size. During training, HIER consumes 38.17 GB of GPU memory, just about 5% more than the 36.21 GB required by Qwen2-VL, thus introducing only minimal additional computational cost. The training time per epoch increases moderately from 1.16 to 1.32 hours due to the additional reasoning complexity introduced by HIER, yet this increase remains well within a reasonable and acceptable range. During inference, HIER attains a latency of 1.04 seconds per sample and a throughput of 432.69 tokens per second, which is sufficiently efficient for deployment in most real-world applications. Besides, the cost of HIER ablations further examine the computational impact of each component. It is observed that removing either the concept clustering module or the self-evolution strategy results in only minor improvements in efficiency, underscoring the lightweight nature of our design and the favorable trade-off between added reasoning capability and computational cost. The added computational cost is primarily concentrated in the relation selection module, yet it introduces only 0.17 million additional parameters and 1.73 GB of extra GPU memory usage. These results show that HIER delivers improved reasoning performance with only a modest increase in computational cost and its training and inference efficiency remain competitive with well-optimized MLLM, underscoring its suitability for deployment in real-world scenarios.

3.2. Hyper Parameters

In this section, we provide the main hyper parameters of HIER and all baselines on MIntRec, MIntRec2.0, and MELD-DA datasets.

3.2.1. HIER

We fine-tune HIER based on the Qwen2-VL[18] architecture using the LoRA[7] technique with a rank of 8, implemented within the LLaMA-Factory[24] framework. Following the pretrained Qwen2-VL encoder, we set the hidden dimension to its default value of 3584. For the number of concept and relation tokens, we set them to 50 and 25, respectively, for the MIntRec dataset. For MIntRec2.0 and MELD-DA, we use 40 and 20 concept tokens, respectively. Instead of applying a threshold-based filtering strategy, we adopt a top-k selection approach to retain the most relevant relation tokens. In the concept clustering module, the number of clustering centers is set based on each dataset's characteristics, aligning with the number of concept tokens. The temperature coefficient is fixed at 1.0 to control the smoothness of the softmax outputs. The parameter α is initialized to 0.5 to balance between soft Spherical KMeans++ clustering and label guidance. The maximum number of iterations is set to 30 to ensure convergence in the clustering process. In the relation extraction module, we configure the MLP layer structure with a hidden dimension of 4096, capturing more complex nonlinear relationships. The training process runs for 5 epochs, using the AdamW [11] optimizer. The learning rate is set to 5e-5 for both the MIntRec and MIntRec2.0 datasets, while a learning rate of 1e-4 is used for MELD-DA. The batch size is fixed at 4, and the learning rate follows a cosine scheduling strategy. To improve training stability and convergence speed, a gradient accumulation step of 8 is applied, with the dropout rate set to 0.1. Besides, bf16 precision is used to enhance computational efficiency. Main training parameters are shown in Table 2.

3.2.2. MISA

For MISA [5], the same configuration is applied across the MIntRec, MIntRec2.0, and MELD-DA datasets. The model uses an LSTM encoder with a hidden size of 256 and a dropout rate of 0.1. Training is conducted for 100 epochs with a batch size of 16, while evaluation and testing use a batch size of 8. All sequences are zero-padded at the end, and no modality alignment is performed. Early stopping is based on the validation F1 score with a patience of 8. Optimization is performed using AdamW with a learning rate of 3e-5, a gradient clipping threshold of -1.0, and a learning rate decay factor $\gamma = 0.5$. A gradient reversal layer is used with a weight of 0.8. For multimodal fusion, the model incorporates weighted losses: reconstruction loss ($w_{\text{recon}} = 0.6$), difference loss ($w_{\text{diff}} = 0.7$), and similarity loss ($w_{\text{sim}} = 0.7$).

3.2.3. MAG-BERT

For MAG-BERT [12], we adopt a training batch size of 16 and an evaluation/testing batch size of 8 on all datatsets. We adopt learning rates of (2e-5, 5e-6, 2e-5) for the MIntRec, MIntRec2.0, and MELD-DA datasets, respectively. The maximum training epoch is set to 100 with early stopping patience of 8 based on the validation metric F1. Zero-padding is applied to the end of sequences, and modality alignment is enabled using the CTC method. The model employs a dropout rate of 0.5 and a beta shift coefficient of 0.005. Weight decay and warmup proportion are fixed at 0.03 and 0.1, respectively.

3.2.4. MulT

For MulT [17], each model is trained for 100 epochs with a training batch size of 16. Evaluation and testing are performed with a batch size of 8, and early stopping is applied with a patience of 8 steps. Zero-padding is applied at the end of each sequence, and modality alignment is disabled. The model uses a shared output dimension of 120 for MIntRec and 80 for both MIntRec2.0 and MELD-DA.

Table 3. Analysis of performance variability on the MIntRec dataset, reported as mean (± standard deviation).

Datasets	MIntRec					
Methods	ACC	F1	P	R	WF1	WP
MISA	72.29(±0.47)	69.24(±0.56)	72.38(±0.55)	73.48(±0.59)	69.32(±0.46)	70.85(±0.52)
MAG-BERT	72.4(±0.56)	68.29(±0.66)	68.87(±0.43)	69.22(±0.45)	72.06(±0.65)	72.04(±0.58)
MulT	72.31(±0.36)	68.97(±0.64)	69.73(±0.57)	68.83(±0.32)	72.07(±0.45)	72.24(±0.62)
TCL-MAP	73.17(±0.63)	68.92(±0.58)	68.9(±0.55)	69.99(±0.7)	72.66(±0.58)	72.97(±0.32)
SDIF-DA	71.64(±0.65)	68.19(±0.39)	69.08(±0.49)	68.3(±0.39)	71.34(±0.56)	71.74(±0.56)
MIntOOD	73.48(±0.5)	70.51(±0.48)	71.19(±0.29)	70.45(±0.51)	72.39(±0.55)	73.24(±0.3)
MVCL-DAF	73.63(±0.54)	70.41(±0.51)	71.07(±0.62)	70.11(±0.39)	73.57(±0.49)	74.31(±0.42)
Qwen2-VL	76.56(±0.68)	74.59(±0.55)	75.85(±0.55)	74.76(±0.5)	76.44(±0.67)	77.19(±0.44)
LLaVA-NeXT	72.65(±0.56)	64.94(±0.54)	66.21(±0.52)	64.65(±0.22)	72.63(±0.27)	73.59(±0.49)
VideoLLaMA2	74.61(±0.59)	71.64(±0.64)	71.82(±0.41)	73.35(±0.34)	74.37(±0.49)	75.43(±0.58)
HIER	80.0(±0.34)	76.91(±0.37)	78.98(±0.51)	77.11(±0.4)	79.59(±0.41)	80.67(±0.44)

Table 4. Analysis of performance variability on the MIntRec2.0 dataset, reported as mean (± standard deviation).

Datasets	MIntRec2.0					
Methods	ACC	F1	P	R	WF1	WP
MISA	55.16(±0.66)	49.51(±0.62)	51.8(±0.71)	49.92(±0.26)	55.05(±0.67)	57.06(±0.45)
MAG-BERT	60.38(±0.53)	54.74(±0.57)	57.51(±0.61)	54.54(±0.41)	59.61(±0.52)	60.0(±0.5)
MulT	60.66(±0.35)	54.12(±0.68)	58.02(±0.56)	53.77(±0.72)	59.55(±0.68)	60.12(±0.66)
TCL-MAP	58.24(±0.7)	52.25(±0.61)	54.28(±0.57)	52.41(±0.39)	57.24(±0.37)	57.55(±0.74)
SDIF-DA	58.06(±0.45)	51.95(±0.55)	53.17(±0.55)	52.16(±0.36)	57.47(±0.33)	57.85(±0.61)
MIntOOD	58.73(±0.54)	52.4(±0.51)	55.48(±0.54)	51.2(±0.36)	58.03(±0.48)	58.34(±0.51)
MVCL-DAF	59.64(±0.39)	53.41(±0.5)	54.9(±0.35)	53.24(±0.36)	58.67(±0.53)	58.57(±0.39)
Qwen2-VL	59.82(±0.43)	47.73(±0.6)	55.23(±0.4)	47.25(±0.35)	58.44(±0.44)	62.51(±0.29)
LLaVA-NeXT	50.61(±0.63)	45.33(±0.4)	50.79(±0.49)	44.14(±0.41)	50.7(±0.17)	54.21(±0.52)
VideoLLaMA2	60.11(±0.44)	49.95(±0.61)	51.58(±0.61)	49.13(±0.39)	59.71(±0.65)	59.9(±0.62)
HIER	64.15(±0.26)	60.31(±0.47)	61.9(±0.42)	59.59(±0.4)	63.79(±0.46)	64.17(±0.46)

The number of Transformer layers and attention heads are set to (8, 8, 2) and (8, 8, 2), respectively, for the MIntRec, MIntRec2.0, and MELD-DA datasets. Learning rates are set to (3e−5, 5e−6, 5e−6) for the three datasets. Dropout rates for attention layers are 0.0 across all datasets, while video and audio modalities use a modality-specific attention dropout of 0.2. Other dropout configurations include ReLU dropout (0.0, 0.3, 0.3), embedding dropout (0.1), residual dropout (0.0), output dropout (0.2), and text dropout (0.4). The model uses attention masking and 1D convolution kernel sizes of (5, 1, 1) for text, video, and audio, respectively. Gradient clipping is applied with a value of 0.5.

3.2.5. TCL-MAP

For TCL-MAP [25], all experiments are trained for 100 epochs, with batch sizes configured as 16 for training and 8 for both evaluation and testing procedures. Zero-padding is applied at the end of each sequence, and modality alignment is enabled. Early stopping is triggered with a patience of 8 based on the validation F1 score. The learning rates

Table 5. Analysis of performance variability on the MELD-DA dataset, reported as mean (± standard deviation).

Datasets	MELD-DA					
Methods	ACC	F1	P	R	WF1	WP
MISA	60.86(±0.31)	49.45(±0.73)	52.7(±0.66)	49.14(±0.36)	58.8(±0.64)	59.55(±0.43)
MAG-BERT	61.08(±0.49)	50.02(±0.61)	52.29(±0.59)	49.85(±0.63)	59.59(±0.39)	59.6(±0.47)
MulT	59.99(±0.5)	50.69(±0.7)	54.83(±0.63)	50.59(±0.53)	58.67(±0.37)	59.39(±0.4)
TCL-MAP	61.63(±0.57)	50.25(±0.52)	53.32(±0.62)	49.64(±0.15)	59.74(±0.42)	60.16(±0.41)
SDIF-DA	60.91(±0.39)	51.46(±0.57)	57.57(±0.5)	50.8(±0.4)	59.58(±0.55)	60.17(±0.32)
MIntOOD	61.58(±0.12)	50.57(±0.57)	54.97(±0.45)	50.63(±0.64)	59.46(±0.4)	60.37(±0.51)
MVCL-DAF	60.78(±0.5)	48.88(±0.49)	54.05(±0.56)	47.73(±0.54)	59.16(±0.6)	59.83(±0.51)
Qwen2-VL	59.71(±0.45)	52.44(±0.44)	55.19(±0.6)	51.93(±0.45)	59.11(±0.34)	59.15(±0.26)
LLaVA-NeXT	51.3(±0.42)	39.87(±0.47)	43.68(±0.28)	38.36(±0.49)	49.77(±0.46)	50.52(±0.54)
VideoLLaMA2	60.81(±0.67)	51.24(±0.58)	56.2(±0.59)	48.79(±0.55)	59.53(±0.6)	59.79(±0.67)
HIER	61.95(±0.32)	54.8(±0.37)	59.41(±0.43)	52.94(±0.45)	60.38(±0.41)	60.44(±0.55)

Table 6. T-test results between HIER and all baseline methods on the MIntRec dataset, reported as t-statistics (p-values), with * indicating significance at p-values less than 0.05.

Datasets	MIntRec					
Methods	ACC	F1	P	R	WF1	WP
MISA	26.32 (4.66e-09)*	22.89 (1.41e-08)*	17.51 (1.15e-07)*	10.21 (7.29e-06)*	33.20 (7.40e-10)*	28.90 (2.22e-09)*
MAG-BERT	23.08 (1.32e-08)*	22.76 (1.47e-08)*	30.21 (1.57e-09)*	26.35 (4.62e-09)*	19.51 (4.94e-08)*	23.86 (1.01e-08)*
MulT	31.00 (1.27e-09)*	21.54 (2.27e-08)*	24.18 (9.13e-09)*	32.50 (8.76e-10)*	24.62 (7.92e-09)*	22.27 (1.75e-08)*
TCL-MAP	18.93 (6.28e-08)*	23.19 (1.27e-08)*	26.87 (3.96e-09)*	17.68 (1.07e-07)*	19.41 (5.16e-08)*	28.54 (2.46e-09)*
SDIF-DA	22.87 (1.42e-08)*	32.30 (9.20e-10)*	27.99 (2.87e-09)*	31.49 (1.13e-09)*	23.80 (1.03e-08)*	25.07 (6.86e-09)*
MIntOOD	21.52 (2.29e-08)*	21.11 (2.67e-08)*	26.42 (4.53e-09)*	20.51 (3.35e-08)*	20.91 (2.87e-08)*	28.06 (2.81e-09)*
MVCL-DAF	19.92 (4.21e-08)*	20.68 (3.14e-08)*	19.61 (4.76e-08)*	24.98 (7.06e-09)*	18.82 (6.56e-08)*	21.01 (2.76e-08)*
Qwen2-VL	9.08 (1.74e-05)*	6.97 (1.16e-04)*	8.32 (3.29e-05)*	7.31 (8.29e-05)*	8.01 (4.31e-05)*	11.22 (3.57e-06)*
LLaVA-NeXT	22.50 (1.61e-08)*	36.55 (3.44e-10)*	35.02 (4.84e-10)*	54.97 (1.33e-11)*	28.25 (2.66e-09)*	21.49 (2.32e-08)*
VideoLLaMA2	15.77 (2.61e-07)*	14.29 (5.60e-07)*	21.78 (2.08e-08)*	14.39 (5.31e-07)*	16.38 (1.94e-07)*	14.49 (5.03e-07)*

are set to (2e-5, 2e-5, 3e-5), and gradient clipping values are (–1.0, –1.0, 0.4). Dropout probabilities for multimodal fusion are (0.5, 0.5, 0.4). A unified SupCon loss is used for contrastive learning, with temperature values of (0.5, 0.5, 0.07). The shared hidden dimension is fixed at 256. All models adopt the CTC-based alignment strategy combined with similarity-based attention. Prompt lengths are set to 3 across all datasets, and label lengths are (4, 4, 3). The transformer prompting module uses 5 attention levels, 8 heads, and consistent dropout settings: 0.1 for attention, 0.2 for embedding, and 0.1 for residual connections. Context encoding with a window size of 3 is used only for MELD-DA.

3.2.6. SDIF-DA

SDIF-DA [8] is trained for 100 epochs with training, evaluation, and test batch sizes set to 16, 8, and 8, respectively. Zero-padding is applied at the end of each sequence, and data alignment is not required. Accuracy is used as the evaluation metric, with early stopping patience set to 8. The feature dimension is fixed at 768 across all datasets. The model employs a two-branch transformer with one level of self-attention and one level of cross-attention. Each self-

Table 7. T-test results between HIER and all baseline methods on the MIntRec2.0 dataset, reported as t-statistics (p-values), with * indicating significance at p-values less than 0.05.

Datasets	MIntRec2.0					
Methods	ACC	F1	P	R	WF1	WP
MISA	25.49 (6.02e-09)*	27.89 (2.95e-09)*	24.45 (8.35e-09)*	40.65 (1.48e-10)*	21.51 (2.30e-08)*	22.03 (1.90e-08)*
MAG-BERT	12.83 (1.29e-06)*	15.09 (3.67e-07)*	11.93 (2.25e-06)*	17.65 (1.08e-07)*	12.11 (2.00e-06)*	12.20 (1.89e-06)*
MulT	16.13 (2.20e-07)*	15.04 (3.78e-07)*	11.15 (3.73e-06)*	14.15 (6.07e-07)*	10.36 (6.51e-06)*	10.01 (8.43e-06)*
TCL-MAP	15.89 (2.46e-07)*	20.86 (2.93e-08)*	21.56 (2.26e-08)*	25.84 (5.39e-09)*	22.32 (1.71e-08)*	15.19 (3.50e-07)*
SDIF-DA	23.67 (1.08e-08)*	23.00 (1.36e-08)*	25.35 (6.29e-09)*	27.68 (3.13e-09)*	22.38 (1.68e-08)*	16.51 (1.83e-07)*
MIntOOD	18.13 (8.79e-08)*	22.77 (1.47e-08)*	18.82 (6.56e-08)*	31.40 (1.15e-09)*	17.41 (1.21e-07)*	16.97 (1.47e-07)*
MVCL-DAF	19.40 (5.18e-08)*	20.12 (3.88e-08)*	25.66 (5.71e-09)*	23.71 (1.07e-08)*	14.61 (4.73e-07)*	18.52 (7.44e-08)*
Qwen2-VL	17.39 (1.22e-07)*	32.90 (7.95e-10)*	23.13 (1.30e-08)*	46.37 (5.17e-11)*	16.82 (1.58e-07)*	6.05 (3.07e-04)*
LLaVA-NeXT	39.97 (1.69e-10)*	48.33 (3.72e-11)*	34.79 (5.10e-10)*	54.25 (1.48e-11)*	53.67 (1.61e-11)*	28.59 (2.42e-09)*
VideoLLaMA2	15.81 (2.56e-07)*	26.87 (3.96e-09)*	28.07 (2.80e-09)*	37.58 (2.76e-10)*	10.30 (6.81e-06)*	11.05 (4.01e-06)*

Table 8. T-test results between HIER and all baseline methods on the MELD-DA dataset, reported as t-statistics (p-values), with * indicating significance at p-values less than 0.05.

Datasets	MELD-DA					
Methods	ACC	F1	P	R	WF1	WP
MISA	4.90 (1.20e-03)*	13.07 (1.11e-06)*	16.99 (1.46e-07)*	13.12 (1.08e-06)*	4.15 (3.22e-03)*	2.54 (3.45e-02)*
MAG-BERT	2.96 (1.82e-02)*	13.47 (8.83e-07)*	19.54 (4.89e-08)*	7.95 (4.59e-05)*	2.81 (2.30e-02)*	2.32 (4.88e-02)*
MulT	6.61 (1.68e-04)*	10.38 (6.43e-06)*	12.03 (2.10e-06)*	6.77 (1.43e-04)*	6.20 (2.60e-04)*	3.07 (1.52e-02)*
TCL-MAP	0.97 (3.59e-01)	14.24 (5.77e-07)*	16.09 (2.24e-07)*	13.92 (6.88e-07)*	2.19 (5.97e-02)	0.81 (4.41e-01)
SDIF-DA	4.14 (3.25e-03)*	9.89 (9.19e-06)*	5.57 (5.29e-04)*	7.10 (1.02e-04)*	2.33 (4.79e-02)*	0.85 (4.22e-01)
MIntOOD	2.15 (6.34e-02)	12.47 (1.60e-06)*	14.22 (5.83e-07)*	5.89 (3.64e-04)*	3.21 (1.25e-02)*	0.19 (8.57e-01)
MVCL-DAF	3.95 (4.24e-03)*	19.36 (5.25e-08)*	15.19 (3.50e-07)*	14.85 (4.17e-07)*	3.37 (9.84e-03)*	1.62 (1.44e-01)
Qwen2-VL	8.14 (3.85e-05)*	8.21 (3.64e-05)*	11.47 (3.03e-06)*	3.17 (1.31e-02)*	4.79 (1.38e-03)*	4.23 (2.90e-03)*
LLaVA-NeXT	40.36 (1.56e-10)*	49.82 (2.92e-11)*	61.07 (5.74e-12)*	43.75 (8.22e-11)*	34.36 (5.63e-10)*	25.64 (5.74e-09)*
VideoLLaMA2	3.08 (1.51e-02)*	10.33 (6.64e-06)*	8.76 (2.26e-05)*	11.71 (2.59e-06)*	2.35 (4.65e-02)*	1.49 (1.74e-01)

attention module uses 8 heads, and each cross-attention module uses 12 heads. The dropout rates for the self and cross branches are set to 0.2 and 0.3, respectively. Weight decay is set to 0.01, and gradient clipping is applied with a maximum norm of 7. The learning rate is 9e-6 for all datasets, with learning rate decay controlled by a patience of 8 and a decay factor of 0.5. Data augmentation is applied only to MIntRec, where augmentation uses a learning rate of 1e-6, batch size of 16, dropout of 0.3, and weight decay of 0.1. The augmentation process runs for 1 epoch and applies no gradient clipping. In contrast, augmentation is disabled for MIntRec2.0 and MELD-DA.

3.2.7. MIntOOD

MIntOOD [21] is trained for up to 100 epochs using an early stopping strategy, with a training batch size of 32 and an evaluation batch size of 16. The base feature dimension is 768 and the warmup proportion is 0.1. The activation uses ReLU with MLP hidden size 256 and dropout 0.1. The minimum and maximum numbers of selected OOD samples are 2 and 3, respectively. The weight dropout is set to 0.5, and the weight hidden dimension is 256. Attention dropout, embedding dropout, and residual dropout are 0.0, 0.0, and 0.2,

respectively. Contrast dropout is 0.1 and number of attention heads is 2. The learning rate is (3e-5, 3e-5, 4e-6), evaluation patience is (8, 8, 3), temperature is (2, 2, 1), alpha is (2, 2, 0.7), MLP dropout is (0.1, 0.1, 0.2), and weight decay is (0.01, 0.01, 0.1) for MIntRec, MIntRec2, and MELD-DA datasets, respectively. Encoder layers for acoustic and visual modalities are set to 1 and 2, respectively.

3.2.8. MVCl-DAF

In all experiments, MVCl-DAF [6] is trained for 100 epochs of training with 16 as the training batch size, while evaluation and testing used a batch size of 8, and early stopping is applied with a patience of 8 steps. The learning rate uses a decay strategy with warmup proportion 0.1 and gradient clipping disabled. Weight decay is set to 0.2 and the shared representation dimension is 256. The model uses the CTC alignment method, MAG epsilon value is 1e-9, and no extra encoder is applied. The evaluation metric differs across datasets, where F1 score is used for MIntRec while accuracy is used for MIntRec2 and MELD-DA. Evaluation patience is (10, 8, 8), and learning rate is (2e-5, 5e-6, 5e-6) for MIntRec, MIntRec2, and MELD-DA, respectively. The contrastive loss is InfoNCE with temperature 0.5. Fusion uses a maximum depth of 5, beta shift 0.006 and dropout probability 0.5.

3.2.9. Qwen2-VL

For Qwen2-VL [18], we adopt the Qwen2-VL-7B-Instruct checkpoint with an encoder hidden size of 3584, and perform LoRA-based fine-tuning on all linear layers under the SFT setting. Training is conducted with a batch size of 2, gradient accumulation steps of 8, a learning rate of 5e-5 following a cosine schedule, for 5 epochs. We use bf16 precision, a warmup ratio of 0.1, and clip gradients with a maximum norm of 1.0. Evaluation is performed every 50 steps, each with a batch size of 2.

3.2.10. LLaVA-NEXT

We utilize the LLaVA-NEXT-Video-7B-hf [10] checkpoint with a hidden size of 4096. Fine-tuning is conducted using LoRA adaptation. The training uses a batch size of 2, gradient accumulation steps of 8, a learning rate of 5e-5 with cosine scheduling for 5 epochs. We apply bf16 precision, a warmup ratio of 0.1, and gradient clipping with a maximum norm of 1.0. Evaluation is performed every 50 steps with a batch size of 2.

3.2.11. VideoLLaMA2

We build on the VideoLLaMA2-7B-16F [3] checkpoint, which use an encoder hidden size of 4096 and a multimodal hidden size of 1024. LoRA is applied to all linear layers. Training is performed with a batch size of 1, gradient accumulation steps of 8, a learning rate of 5e-5 using cosine scheduling, and FP16 precision over 5 epochs except for 3 for mintrec. A warmup ratio of 0.1 and a max gradient norm of 1.0 are used. Evaluation occurs every 50 steps with a batch size of 1.

3.3. Performance Variability and Significance Test

To evaluate the robustness and significance of HIER, we conduct a comprehensive performance variability and significance test with all baselines across three benchmark datasets including MIntRec, MIntRec2.0, and MELD-DA. The results are reported as mean (± standard deviation) in Table 3, Table 4, and Table 5 for each dataset while t-statistics (p-values) in Table 6, Table 7, and Table 8.

For performance variability, HIER consistently achieves the best overall performance across all three datasets and six metrics, while also demonstrating lower standard deviations, indicating stable and reliable predictions. For instance, on the MIntRec dataset, HIER achieves an ACC of 80.00% (±0.34%), substantially outperforming all baselines, with a standard deviation less than half that of most compared methods. A similar pattern is observed for other key metrics such as F1, P, and WF1, all of which show both higher mean values and lower variances compared to baselines. This trend extends to the MIntRec2.0 and MELD-DA datasets as well, where HIER continues to maintain top-tier performance with minimal variability. The combination of high average scores and tight performance bounds underlines HIER's robustness and reliability in diverse scenarios, suggesting its strong generalization capability.

For the significance test, results on the MIntRec dataset reveal that all p-values fall well below the conventional 0.05 threshold, often reaching exceedingly small magnitudes accompanied by large t-statistics, strongly confirming the statistical confidence in HIER's superiority. On the more challenging MIntRec2.0 dataset, HIER's advantages become even more pronounced. All t-tests across all baselines and evaluation metrics yield statistically significant differences against powerful LLM-based models such as Qwen2-VL, where HIER achieves a significant improvement in ACC with a t-statistic of 17.39 and a p-value of 1.22e-07, and LLaVA-NeXT, where it achieves a notable advantage in F1 with a t-statistic of 48.33 and a p-value of 3.72e-11. On the MELD-DA dataset, although the overall performance gap between models narrows, primarily due to the inherent ambiguity of dialogue act categories, HIER continues to achieve statistically significant improvements over most baselines, particularly in core metrics such as F1, WF1, and WP. Notably, in comparison with the strong VideoLLaMA2 baseline, HIER demonstrates clear statistical superiority across multiple aspects. For example, it yields substantial gains in P, with a t-statistic of 8.76 and a p-value of 2.62e-05, as well as in R, with a t-statistic of 11.71 and a p-value of 2.59e-06. These results further underscore HIER's robustness and consistent effectiveness in tackling complex

Table 9. Prompt sensitivity analysis of HIER on MIntRec2.0, examining the impact of three prompt variations including Lexical Substitution, Semantic Rephrasing and Formatting Changes.

Variants	ACC	F1	P	R	WF1	WP
Lexical Substitution	62.72	58.88	61.02	58.20	62.24	63.01
Semantic Rephrasing	63.16	59.34	61.46	58.39	62.05	63.16
Formatting Changes	63.04	60.11	60.98	59.07	61.94	62.79
HIER	**64.15**	**60.31**	**61.90**	**59.59**	**63.79**	**64.17**

Table 10. Comparison between HIER and CoT-enhanced MLLMs on the MIntRec2.0 dataset, including Qwen2-VL + ***CoT***, LLaVA-NeXT+ ***CoT***, and VideoLLaMA2+ ***CoT***.

Methods	ACC	F1	P	R	WF1	WP
Qwen2-VL	59.82	47.73	55.23	47.25	58.44	62.51
+CoT	62.57	49.66	51.72	48.86	62.55	63.57
LLaVA-NeXT	50.61	45.33	50.79	44.14	50.70	54.21
+CoT	58.53	51.62	58.26	51.95	56.93	59.76
VideoLLaMA2	60.11	49.95	51.58	49.13	59.71	59.90
+CoT	61.55	53.57	54.48	52.92	61.28	61.09
HIER	**64.15**	**60.31**	**61.90**	**59.59**	**63.79**	**64.17**

Table 11. Results of HIER using latest backbones on MELD-DA.

Methods	ACC	F1	P	R	WF1	WP
Qwen2.5-VL	59.81	46.57	47.73	46.39	59.15	58.92
+HIER	**62.04**	**53.74**	**59.14**	**53.87**	**60.79**	**60.05**
Qwen3-VL	59.91	52.57	54.94	51.99	59.34	59.30
+HIER	**63.93**	**54.64**	**59.76**	**53.45**	**62.29**	**61.13**

multimodal intent recognition tasks, owing to the hierarchical semantics and evolutionary reasoning paradigm.

4. Additional Discussions

4.1. Prompt Sensitivity Analysis

To further assess the robustness of HIER against prompt variability, we conduct a prompt sensitivity analysis on the MIntRec2.0 dataset. Following prior work on prompt sensitivity [13], we consider three representative types of variations. First, Lexical Substitution replaces key terms in the original prompt with synonyms to evaluate the model's sensitivity to word choice. Second, Semantic Rephrasing reformulates the prompt using alternative sentence structures while preserving its original meaning. Third, Formatting Changes adjusts the visual structure of the prompt by modifying elements such as line breaks and indentation without altering the prompt's content.

Table 12. Comparison with closed-source models on MIntRec2.0.

Methods	ACC	F1	P	R	WF1	WP
Gemini-3 Pro	51.65	47.25	50.39	49.51	52.45	58.11
GPT-4o	42.32	37.49	42.38	42.01	43.60	52.98
HIER	**64.15**	**60.31**	**61.90**	**59.59**	**63.79**	**64.17**

Experimental results are summarized in Table 9, revealing that HIER consistently maintains high stability across diverse prompt variations. In the Lexical Substitution setting, performance degradation remains within 1.5% across most evaluation metrics, representing only a minor fluctuation compared to the substantial improvements HIER delivers over the baseline. Furthermore, Semantic Rephrasing and Formatting Changes lead to even smaller deviations from the original prompt, with the smallest drop reaching just 0.2%. These results clearly demonstrate that HIER's hierarchical reasoning remains focused on semantic intent despite surface-level changes in phrasing or structure. This resilience confirms HIER's robustness and highlights its practical applicability and generalizability in real-world scenarios where prompts may vary in form and style.

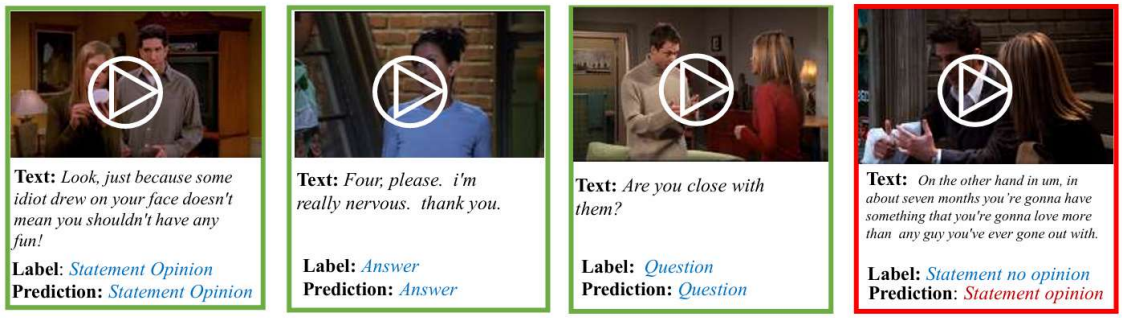
4.2. Comparison with CoT-enhanced MLLMs

To comprehensively validate the effectiveness of HIER, we compare it against three leading MLLMs and their CoT-enhanced variations on the MIntRec2.0 dataset. The selected MLLMs include Qwen2-VL, LLaVA-NeXT, and VideoLLaMA2. The CoT template follows the design described in Section 2.1, with the hierarchical feature tokens removed. All experimental conditions are kept consistent with those used in the baseline MLLM evaluations. The results are exhibited in Table 10.

First, we can observe that HIER consistently outperforms all CoT-enhanced baselines, achieving gains exceeding 1.5% across all metrics except for the WP. Besides, HIER delivers remarkable improvements over Qwen2-VL + ***CoT***, surpassing it by more than 10% in both F1-score and precision, underscoring the advantage of the hierarchical semantic representation. Second, all MLLMs exhibit notable performance gains after being augmented with the CoT strategy. In particular, LLaVA-NeXT shows improvements exceeding 5% across all evaluation metrics. These results highlight the effectiveness and soundness of our CoT design, demonstrating its ability to enhance the reasoning capabilities of MLLMs when handling complex semantics. Together, these findings confirm that while CoT prompting provides a solid foundation for reasoning, HIER further elevates performance through explicit semantic hierarchies.

(a) Examples from the MIntRec dataset.

(b) Examples from the MIntRec2.0 dataset.



(c) Examples from the MELD-DA dataset.

Figure 1. Representative examples of correctly and incorrectly predicted cases from three datasets: MIntRec (a), MIntRec2.0 (b), and MELD-DA (c).

4.3. Generalization to Latest Backbones

To evaluate the generalization capability, we integrate it with the latest multimodal backbones and conduct experiments on the MELD-DA dataset. Specifically, we adopt Qwen2.5-VL and Qwen3-VL as representative MLLMs and maintain identical experimental settings to ensure fair comparisons. The results are summarized in Table 11.

It is evident that HIER consistently improves performance across all evaluation metrics when built upon the latest backbones. For Qwen2.5-VL, incorporating HIER yields notable gains, with F1 increasing by 7.17% and P rising by 11.41%, alongside consistent improvements in other metrics. Similar improvements are observed for Qwen3-VL, where HIER boosts ACC from 59.91% to 63.93% and F1 from 52.57% to 54.64%. Importantly, the consistent improvements indicate that HIER maintains strong adaptability to different backbones, confirming that the hierarchical semantic modeling and evolutionary reasoning mechanisms provide a general and scalable enhancement for complex multimodal intent understanding.

4.4. Comparison with Closed-source Models

To assess the competitiveness of HIER, we compare it with two leading closed-source models, GPT-4o and Gemini-3 Pro, on the MIntRec2.0 dataset. The closed-source models are evaluated under few-shot setting with detailed CoT. The experimental results are shown in Table 12.

It can be observed that HIER achieves substantial improvements over both closed-source baselines across all evaluation metrics. Specifically, HIER surpasses Gemini-

3 Pro by more than 13% in ACC and 13% in F1, and exceeds GPT-4o by over 20% in ACC and nearly 23% in F1, which highlights the effectiveness of explicit hierarchical semantic modeling and evolutionary reasoning. Similar advantages are observed in P, R, and WF1 with clear gains over 10%, further confirming the robustness of the improvements. These results demonstrate that even when compared with powerful closed-source MLLMs equipped with CoT prompting, HIER maintains a clear and consistent lead, suggesting that structured semantic abstraction can provide strong reasoning benefits.

4.5. Case Study

We conduct case studies on the MIntRec, MIntRec2, and MELD-DA datasets, selecting three correctly predicted representative instances and one challenging mispredicted case per dataset, as illustrated in Figure 1, which aims to uncover the operational patterns and failure modes of our method when faced with varying semantic complexities. Besides, to illustrate the concrete meaning of the semantic hierarchy, we select a representative example from the MIntRec2.0 dataset and present the intermediate CoT outputs produced by the trained HIER model, as shown in Figure 2.

We begin our analysis with the correctly predicted examples, which consistently exhibit clearly identifiable semantic concepts and coherent cross-modal interrelations. In the first case from MIntRec, the utterance follows a typical command structure, and the accompanying video reinforces this interpretation through a workplace setting and recognizable character roles. The alignment between behavioral cues in the text and the situational context provided by the video allows the model to accurately infer the communicative intent. In the third example from MIntRec2.0, although discourse markers such as "Yeah", "well", and "so" might mislead the model toward predicting intents like *Agree* or *Explain*, the speaker's tilted head and subtly negative facial expression align more closely with textual negation cues like "not", enabling the model to correctly resolve the ambiguity. In the MELD-DA dataset, similar cross-modal dependencies remain evident. For instance, in the third example, the speaker's facial expressions and gestures, along with the interrogative structure of the text, provide consistent semantic signals that help guide the model to the correct intent classification. These successes underscore the effectiveness of HIER's hierarchical semantic representation and evolutionary reasoning strategy, which together enable the model to identify key semantic cues, reason over their structural dependencies, and deliver robust intent predictions in complex, multimodal scenarios. Failure cases from the MIntRec and MELD-DA dataset highlight a recurring challenge in which the model is misled by surface-level semantic cues, such as overt action phrases like "gonna check" or emotionally intense statements like "I love him more than any guy". While these expressions carry strong semantic information, they tend to dominate the model's focus and obscure the deeper communicative semantic salience interference frequently arises in complex intent categories such as *Taunt* and *Joke*, reflecting the difficulty of disentangling prominent linguistic elements from the speaker's underlying communicative goals, remaining a fundamental challenge for achieving robust and context-aware multimodal intent recognition. In contrast, the failure case from MIntRec2.0 reveals a distinct limitation stemming from the absence of discourse-level context. The model's misclassification of a semantically ambiguous utterance as an *Agree* intent demonstrates the weakness in capturing discourse dependencies that are crucial for accurate pragmatic inference. Despite these challenges, HIER still achieves strong performance, attaining high accuracy across multiple datasets. These insights not only affirm the effectiveness of HIER in real-world multimodal scenarios but also point to promising avenues for future research.

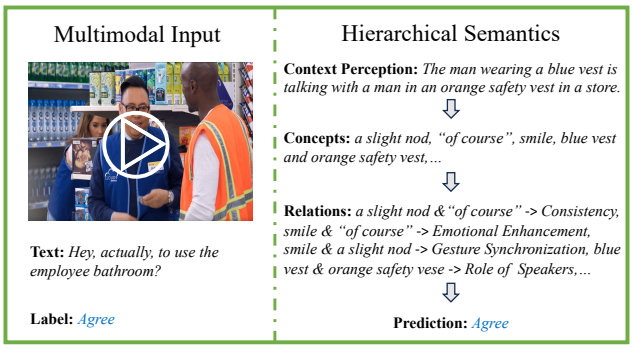


Figure 2. A detailed example showcasing the semantic hierarchy.

Then we present an illustrative example in Figure 2 to give a deeper insight of semantic hierarchy. In the initial step, the model accurately describes the visual scene *"The man wearing a blue vest is talking with a man in an orange safety vest in a store."* as context perception. Subsequently, in the concept extraction stage, HIER identifies semantically relevant cues across multiple modalities such as the gesture *a slight nod*, the phrase *"of course"*, the facial expression *smile* and visual attributes *blue vest* and *orange safety vest*. In the final relational reasoning step, HIER successfully infers the semantic proximity between these concepts. For example, it associates the gesture *a slight nod* with the phrase *"of course"* to establish a relation of semantic consistency, links *smile* and *"of course"* to indicate emotional enhancement, and connects *smile* with *a slight nod* to infer gesture synchronization. Furthermore, the visual elements *blue vest* and *orange safety vest* are integrated to derive the role of speakers. By leveraging these multimodal relations, the model concludes that the underlying communicative intent is **Agree**. This example showcases HIER's ability to perform structured, interpretable reasoning across hierarchical semantic levels.

References

[1] Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025. Accessed: 2025-04-08. 1

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 1

[3] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 7

[4] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms, 2024. 1

[5] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131, 2020. 3

[6] Bo Hu, Kai Zhang, Yanghai Zhang, and Yuyang Ye. Adaptive multimodal fusion: Dynamic attention allocation for intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17267–17275, 2025. 7

[7] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3

[8] 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)*, pages 10206–10210. IEEE, 2024. 5

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 1

[10] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 7

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 3

[12] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for computational linguistics. Meeting*, page 2359, 2020. 3

[13] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024. 8

[14] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 1

[15] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: solving ai tasks with chatgpt and its friends in hugging face. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 1

[16] Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, page 887–890, New York, NY, USA, 2024. Association for Computing Machinery. 2

[17] 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 *Proceedings of the conference. Association for computational linguistics. Meeting*, page 6558, 2019. 3

[18] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 3, 7

[19] Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators, 2025. 1

[20] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 1

[21] 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*, 2025. 6

[22] Wenqiao Zhang, Tianwei Lin, Jiang Liu, Haoyuan Li, Fangxun Shu, Wanggui He, Zhelun Yu, Lei Zhang, Zheqi Lv, Hao Jiang, Juncheng Li, Siliang Tang, and Yueting Zhuang. HyperLLaVA: Dynamic visual and language expert tuning for multimodal large language models, 2024. 1

[23] Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 1

[24] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3

[25] 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. In *Proceedings of the AAAI conference on artificial intelligence*, pages 17114–17122, 2024. 4