SEQUENCE MATTERS: A CONTROLLED STUDY OF MODALITY INJECTION ORDER IN MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) integrate heterogeneous inputs such as text, vision and audio and are increasingly fine-tuned as pretrained backbones across various downstream classification tasks. While recent studies suggest that the ordering of multimodal inputs may affect model behavior, existing work remains largely limited to isolated settings, focusing on a single dataset, task or model. We introduce a Modality Order-aware Controlled Architecture for MLLMs namely *MOCA-MLLM*, that enables controlled multimodal scheduling during decoder processing. This work is positioned as a controlled empirical study rather than a mechanistic interpretability analysis. Building upon this framework, we conduct a large-scale empirical study on the impact of modality injection order during fine-tuning, across four MLLMs, six downstream tasks, twelve multimodal datasets, three modalities and six modality orders. Our results reveal consistent and task-dependent performance patterns with regard to modality order, demonstrating that the choice of injection sequence can substantially influence fine-tuning outcomes under the present controlled setting. These findings reveal modality order as an empirically consequential yet underexplored design variable in downstream multimodal adaptation and provide practical guidance for future MLLM fine-tuning. Furthermore, comparisons with an All-at-Input baseline reveal that simply increasing the processing depth of all modalities does not consistently improve performance, suggesting that modality scheduling merits careful consideration in downstream adaptation.

1 INTRODUCTION

Recent years have witnessed rapid progress in Large Language Models (LLMs), which have demonstrated remarkable capabilities in representation learning and reasoning across a wide range of natural language processing tasks Chang et al. (2024); Minaee et al. (2025); Yang et al. (2025). Building upon this success, increased attention has been devoted to extending LLMs toward Multimodal LLMs (MLLMs), enabling them to process and reason over heterogeneous inputs such as text, vision and audio Yin et al. (2024); Jin et al. (2025).

Multimodal understanding has become particularly important for downstream tasks including sentiment analysis, medical diagnosis and several multimodal classification/generation, where relying on a single modality often leads to incomplete or biased representations Zheng & Zhang (2025); Guo et al. (2023); Zhan et al. (2025). MLLMs adopt a pretrained LLM as a unified backbone and inject multimodal information in the form of token sequences Abdin et al. (2024). This architecture enables deep cross-modal reasoning and a new degree of freedom with regard to the order and timing with which different modalities are incorporated into these models. Since transformer-based language models are inherently sequential Vaswani et al. (2017); Chen et al. (2024a); Sajun et al. (2024), their attention patterns and intermediate representations are explicitly conditioned on token positions and contextual ordering. Consequently, altering the sequence of modality injection may influence how multimodal information is integrated and represented within these models.

Recent studies have begun to report that MLLMs can exhibit notable sensitivity to the ordering of multimodal inputs, with downstream performance varying substantially under different modality

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.

sequences Tomar et al. (2023); Tan et al. (2024). However, existing investigations remain largely fragmented and limited in scope. Most prior studies examine ordering effects under highly constrained settings, such as a single dataset, a specific task, or a particular model, and primarily focus on prompt-level or inference-time sequencing rather than model adaptation during training Tomar et al. (2023); Tan et al. (2024); Wardle & Sušnjak (2025); Guan et al. (2025). Consequently, it remains unclear whether the observed modality order effects reflect a general property of MLLMs or merely dataset- or model-specific artifacts that arise under particular experimental conditions.

Research Questions. Motivated by these observations, we conduct a systematic investigation into the role of modality injection order in fine-tuning pretrained MLLMs for downstream tasks. Specifically, we seek to answer the following questions: **(i)** Does the order of modality injection lead to consistent differences in downstream task performance during fine-tuning? **(ii)** Are certain modality orders more suitable for particular tasks, datasets or model architectures? **(iii)** Does increasing modality processing depth necessarily lead to improved multimodal performance or does the timing of modality integration play a more critical role?

To this end, we propose a Modality Order-aware Controlled Architecture for MLLMs, namely *MOCA-MLLM* and perform a large-scale empirical study across diverse backbone architectures, downstream tasks and datasets and modality combinations. This work is positioned as a controlled empirical study: our goal is to identify robust empirical trends in downstream adaptation behavior and we intentionally avoid over-interpreting the underlying mechanisms, which we view as a valuable direction for future work. To further disentangle the effect of processing depth from modality scheduling, we additionally introduce an *All-at-Input* baseline, where all modalities are injected at the decoder input and processed through the full depth of the transformer. This baseline, which approximates omni-modal simultaneous exposure, serves as a strong and practically relevant reference point; outperforming it demonstrates that staged modality exposure can yield benefits beyond simply maximizing processing depth.

The **contributions** of this paper are summarized as follows: **(i)** We present the first controlled, large-scale empirical study of modality injection order during MLLM fine-tuning across multiple backbone architectures, downstream tasks and modality combinations. We design *MOCA-MLLM*, a lightweight order-aware framework compatible with existing decoder architectures that enables systematic investigation of modality scheduling effects, as illustrated in Figure 1. We additionally introduce an *All-at-Input* baseline to disentangle the impact of modality order from processing depth and to study whether downstream adaptation benefits from staged modality exposure relative to simultaneous exposure. **(ii)** We conduct a comprehensive empirical study across 4 backbone MLLMs, 6 downstream tasks, 12 multimodal datasets, 3 modalities and all 6 possible modality orders, providing one of the most extensive evaluations of modality scheduling effects to date. **(iii)** We provide consistent empirical evidence that certain staged modality schedules outperform simultaneous All-at-Input exposure during downstream adaptation and that task-dependent modality order preferences are reproducible across backbone architectures and datasets. These findings offer practical guidance without requiring theoretical claims about intrinsic modality properties.

2 Related Work

Multimodal LLMs. Multimodal Learning aims to enable models to jointly reason over heterogeneous modalities such as text, vision and audio Yin et al. (2024); Jin et al. (2025). A line of work demonstrates that frozen or lightly adapted LLMs can serve as central reasoning backbones by treating non-text modalities as sequences or prefixes, enabling multimodal generation and reasoning without retraining the entire model Tsimpoukelli et al. (2021); Alayrac et al. (2022); Li et al. (2023); Liu et al. (2023). More recent work further extends this paradigm by scaling multimodal reasoning capabilities across diverse modalities and investigating sparse or decoupled transformer architectures, forming the new generation of general-purpose MLLMs Yang et al. (2023); Chen et al. (2024b); Deitke et al. (2024); Zhou et al. (2025); Walsh et al. (2025); Liang et al. (2025).

Modal Order Sensitivity. Beyond multimodal mechanisms, recent studies suggest that neural models can be sensitive to the sequence and ordering of their inputs. In language understanding, robustness evaluations reveal that altering the structure or ordering of premises can substantially affect inference behavior and generalization Nie et al. (2020); Ribeiro et al. (2020); Gardner et al.

(2020). With regard to in-context learning, multiple studies have shown that the selection, ordering and formatting of demonstrations play a critical role in shaping model performance, indicating that input organization noticeably influences model behavior Min et al. (2022); Liu et al. (2022); Zhao et al. (2021); Lu et al. (2022). Extending these observations to multimodal settings, several task-specific studies explicitly incorporate modality order or structured modality dependencies and report improved performance on multimodal classification or reasoning tasks Tomar et al. (2023); Tan et al. (2024); Wardle & Sušnjak (2025); Guan et al. (2025). More recently, Gupta et al. Gupta & Gandelsman (2026) investigated modality-order consistency at test time through adaptive consistency optimization, providing complementary evidence that modality order substantially influences multimodal reasoning.

Nevertheless, a systematic study that isolates and evaluates modality injection order during fine-tuning, across diverse models, tasks, datasets and modality combinations, remains under-explored.

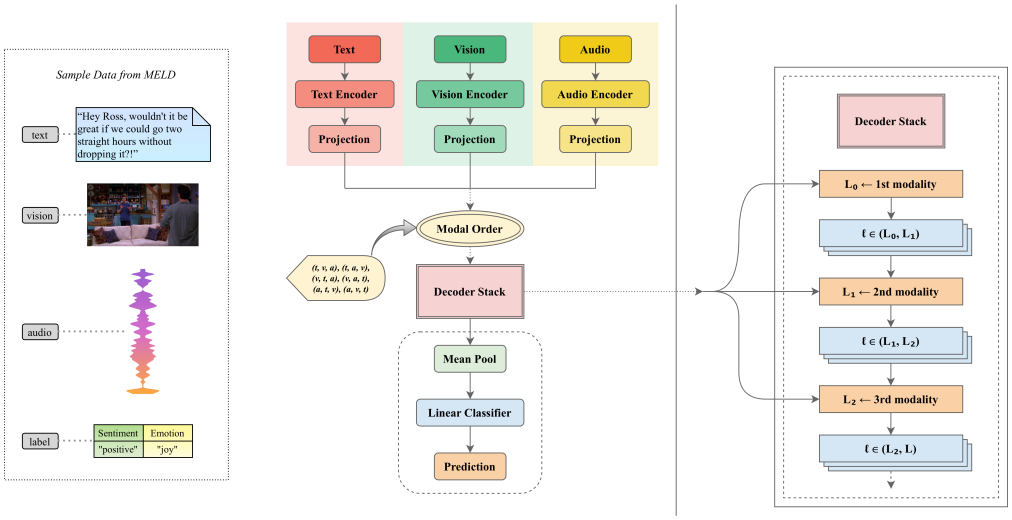


Figure 1: Overall architecture of the proposed framework *MOCA-MLLM*. Modalities are injected into the decoder at different depths following a controlled order-wise strategy.

3 METHODOLOGY

As illustrated in Figure 1, MOCA-MLLM integrates modalities via external frozen pretrained encoders whose projected representations are inserted into the LLM hidden space at designated decoder layers. No decoder redesign is performed and no custom fusion module replaces backbone behavior, making this an architecture-consistent abstraction compatible with practical MLLMs such as Qwen, LLaVA, Phi, and Gemma variants.

Unified Modality Embedding Pipeline. A critical prerequisite for this study is ensuring that modality representations reside in a comparable semantic space. Since pretrained LLMs are typically equipped with modality-specific encoding or fusion mechanisms, directly applying their internal encoders would introduce uncontrollable bias. Therefore, we employ a unified processing pipeline with three frozen pretrained encoders to process extracted text, vision and audio inputs.

For modality $m \in \{\text{text}, \text{vision}, \text{audio}\}$, the raw input is first mapped to a modality-specific token sequence of hidden representations, $H_m \in \mathbb{R}^{T_m \times d_m}$, where T_m denotes the sequence length and d_m is the feature dimensionality. The sequence length T_m is determined by the inherent structure of the modality and dataset-specific preprocessing and is not forced to be aligned across modalities.

We denote the projected modality embedding as $\tilde{H}_m \in \mathbb{R}^{T_m \times d_{\text{LLM}}}$, obtained via offline linear projection layers applied at the token level t:

$$\tilde{H}_m^{(t)} = W_m H_m^{(t)} + b_m, \; t \in \{1, \ldots, T_m\},$$

where $W_m \in \mathbb{R}^{d_{\text{LLM}} \times d_m}$ and $b_m \in \mathbb{R}^{d_{\text{LLM}}}$ are modality-specific parameters, d_{LLM} denotes the hidden dimensionality of the language model decoder and $H_m^{(t)} \in \mathbb{R}^{d_m}$ and $\tilde{H}_m^{(t)} \in \mathbb{R}^{d_{\text{LLM}}}$ denote the input and projected representations of the t-th token, respectively.

All modality encoders are kept frozen throughout training. The modality-specific projection layers, however, are unfrozen and trained together with the designated decoder layers and the classification head, allowing the model to learn a task-adapted mapping from each modality's encoder space into the LLM hidden space. This design isolates the effect of modality injection order from encoder-side representation learning, while still permitting projection-level adaptation.

Modal Order Permutations. Modality order in this work refers to sequential injection at different decoder depths, rather than token-level permutations at a single layer. We systematically evaluate the effect by considering all possible permutations of the three available modalities

$$\Pi = \{(t, v, a), (t, a, v), (v, t, a), \\ (v, a, t), (a, t, v), (a, v, t)\}.$$

where t, v, and a denote text, vision and audio, respectively. In later sections, we use the shorthand *x-y-z* to denote the modality order (x, y, z).

Controlled Injection Strategy. Modality embeddings are injected at three decoder layers indexed as L_0, L_1 and L_2, corresponding to the input layer and two deeper layers. Let $L_0 = 0$, $L_1 = \lceil L/3 \rceil$ and $L_2 = \lceil 2L/3 \rceil$, where L denotes the total number of decoder layers. This setting ensures distinct fusion depths while allocating each modality a comparable number of engaged decoder layers.

At each insertion layer, multimodal fusion is performed by directly concatenating the projected modality representations along the sequence dimension, prior to the forward computation of the insertion layer itself. Let $X^{(\ell)} \in \mathbb{R}^{T^{(\ell)} \times d_{\text{LLM}}}$ represent the decoder hidden states at layer ℓ. Given a modal order as $\pi = (m_1, m_2, m_3) \in \Pi$, the very first layer is initialized as $X^{(L_0)} = \tilde{H}_{m_1}$. At the two insertion layers L_1 and L_2, the new modality is injected by concatenating to the hidden states:

$$X^{(L_1)} = \left[X^{(L_1)}; \tilde{H}_{m_2} \right], X^{(L_2)} = \left[X^{(L_2)}; \tilde{H}_{m_3} \right],$$

where $[\cdot;\cdot]$ denotes concatenation along the sequence dimension.

Between insertion layers, the decoder processes the hidden states in a standard layer-wise manner. Let $f_\ell(\cdot)$ denote the ℓ-th decoder block and the hidden states follow $X^{(\ell+1)} = f_\ell\big(X^{(\ell)}\big)$, while the dimensionalities evolve as

$$X^{(\ell)} \in \mathbb{R}^{T_{m_1} \times d_{\text{LLM}}} \text{ for } \ell \in [L_0, L_1), \\ X^{(\ell)} \in \mathbb{R}^{(T_{m_1}+T_{m_2}) \times d_{\text{LLM}}} \text{ for } \ell \in [L_1, L_2), \\ X^{(\ell)} \in \mathbb{R}^{(T_{m_1}+T_{m_2}+T_{m_3}) \times d_{\text{LLM}}} \text{ for } \ell \in [L_2, L).$$

The base model is initially fully frozen. Subsequently, an equal number of decoder layers following each injection block, together with the three modality-specific projection layers and the classification head, are unfrozen for fine-tuning.

Loss and Optimization. All tasks are formulated as supervised classification problems and optimized using the cross-entropy loss, while for highly imbalanced datasets we apply weighted cross-entropy instead. Meanwhile, model parameters are optimized using AdamW with weight decay. We adopt a partitioned learning-rate strategy, assigning distinct learning rates to the unfrozen decoder layers and the classification head. We prioritize stable optimization and competitive performance to ensure the validity of our empirical comparisons.

4 Experimental Setup

Models, Encoders, Tasks, and Datasets. We evaluate four SOTA MLLMs, serving as the backbone for our progressive fusion experiments: *Phi-3.5-Vision-Instruct* Abdin et al. (2024), *Qwen2.5-Omni-7B* Xu et al. (2025), *LLaVA-1.5-7B* Liu et al. (2023), *Gemma-3-4B-IT* Team et al. (2025). We

utilize consistent encoders: *MiniLM-L6-H384-uncased* for text Wang et al. (2020), *CLIP-ViT-Base-Patch32* for vision Radford et al. (2021) and *Wav2Vec2-Base-960h* for audio Baevski et al. (2020). Our evaluation covers six task categories and twelve multimodal datasets to ensure the generalizability of our findings, as shown in Table 1.

Training Protocol. The unfreezing strategy is strictly controlled. The primary variable investigated is the modality injection order and thus, we identify the optimal parameter settings, including batch size, learning rate and weight decay, prior to final evaluation.

Evaluation Metrics. We report downstream performance using standard classification metrics: *Accuracy*, *Precision*, *Recall* and *F1-Score* (all macro-averaged). To mitigate the variance inherent in multimodal training, we conduct 20 repeated experiments for each model setting and consider mean performance over 20 independent runs as the primary evaluation metric. Additional statistical results and analysis including standard deviation and t-tests can be found in the supplementary material.

All-at-Input Baseline. In addition to sequential modality injection, we introduce an *All-at-Input* baseline to examine whether the observed performance differences are primarily caused by processing depth. In this baseline setting, the projected representations of all modalities are concatenated with the text tokens and injected at the input layer of the decoder. Consequently, all modalities are processed through the full depth of the transformer.

If deeper processing alone were the dominant factor, the All-at-Input setting would be expected to outperform sequential injection across tasks.

5 Results and Analysis

5.1 Sequential Injection Analysis

Table 1 reports the full accuracy and F1-score results for each configuration. Table 2 further summarizes the observed modality-order performance patterns using averaged accuracy values with relative color-based highlighting. Additional statistical analyses, including mean$\pm$standard deviation results and representative t-test comparisons, are provided in supplementary material.

Sentiment Analysis. For CMU-MOSI Zadeh et al. (2016), the *v-t-a* and *a-v-t* sequences consistently achieve the highest or second-highest accuracy; for MELD Poria et al. (2019) (Figure 2a), text-first orders (*t-v-a* and *t-a-v*) consistently yield the lowest accuracy across all four backbone models, dominated by vision-first and audio-first strategies. This reflects an empirically observed tendency in which, under the present controlled setting, early textual injection appears to bias the model toward surface-level semantic cues, potentially constraining integration of the paralinguistic and visual affective signals that convey sentiment in conversational audiovisual data.

Across both datasets and all four backbone models, the cross-dataset agreement is notable: audio and vision early scheduling consistently outperforms text-first in three-class sentiment classification. The *v-t-a* order performs particularly strongly on CMU-MOSI, which contains dyadic interview recordings where prosodic and facial cues are highly correlated with sentiment intensity. For MELD, which involves multi-party conversational data with faster turn-taking, audio-first configurations are more uniformly preferred. This dataset-level divergence in the best-performing specific order, alongside consistent agreement that text-first is least favorable, suggests the pattern is a task-level empirical regularity rather than a dataset artifact.

Emotion Recognition. In three out of four models, text-first orders (*t-v-a* and *t-a-v*) generally yield the lowest performance, while audio-first or delayed configurations (*v-a-t*, *a-t-v*, *a-v-t*) achieve the strongest results. This pattern holds across both 7-class MELD (Figure 2b) and 5-class IEMOCAP Busso et al. (2008), suggesting a task-level preference rather than a dataset-specific artifact. LLaVA-1.5 exhibits a more model-dependent pattern: vision-first orders consistently under-perform while vision-middle orders lead. Overall, audio-first injection yields the strongest results, consistent with the hypothesis that paralinguistic cues contribute more effectively when allocated greater decoder processing depth.

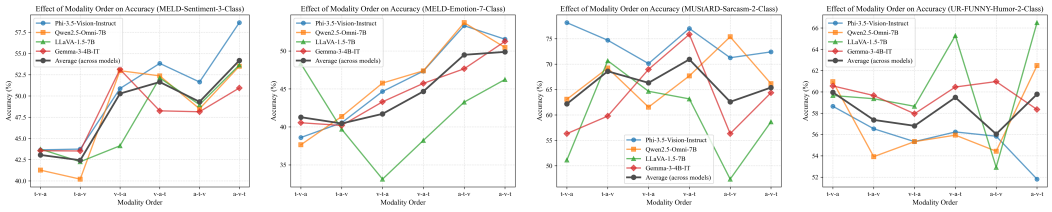
Table 1: Performance comparison under different modality injection orders across six tasks (number of classes indicated), twelve datasets, and four backbone models. Each cell reports Accuracy / Macro-F1 (in %). For each model and dataset, **Bold** (Plain) denotes the highest (second highest) accuracy, and **Bold** (Plain) denotes the lowest (second lowest) accuracy amongst all modality orders.

Order	Phi	Qwen	LLaVa	Gemma	Phi	Qwen	LLaVA	Gemma
Sentiment Analysis								
	MELD (3) Poria et al. (2019)				**CMU-MOSI (3)** Zadeh et al. (2016)			
t-v-a	43.65/29.52	41.26/26.30	43.74/28.69	43.55/25.63	50.20/45.74	**46.24**/42.46	50.55/30.10	52.33/40.58
t-a-v	**43.74**/29.21	**40.20**/26.12	**42.25**/27.84	**43.52**/26.12	**46.97**/41.67	46.56/44.06	50.39/35.64	52.00/40.32
v-t-a	50.88/35.87	52.98/36.19	44.13/26.03	**53.14**/37.61	51.98/48.91	48.63/42.22	**55.10**/42.49	54.13/44.87
v-a-t	53.84/39.34	52.37/37.80	52.12/33.04	48.25/31.93	51.25/45.94	46.30/42.27	53.00/40.71	**49.83**/39.36
a-t-v	51.65/35.58	48.49/34.01	48.98/33.60	48.14/33.59	50.50/46.91	46.85/42.18	**50.38**/37.99	51.01/39.91
a-v-t	**58.64**/37.82	**53.50**/36.36	**53.65**/38.04	50.94/35.67	**54.08**/46.46	**48.94**/46.41	54.08/39.89	**54.24**/43.35
Emotion Recognition								
	MELD (7)				**IEMOCAP (5)** Busso et al. (2008)			
t-v-a	**38.60**/25.58	**37.64**/23.05	**48.32**/34.03	40.54/25.04	**43.24**/44.23	**43.65**/43.78	**59.46**/59.49	41.08/40.62
t-a-v	40.54/26.68	41.37/28.26	39.69/25.12	**40.19**/26.21	44.73/45.50	45.95/45.84	50.54/50.88	48.11/48.30
v-t-a	44.66/31.91	45.77/32.43	**33.12**/18.93	43.30/29.09	55.81/56.31	51.47/56.49	**36.62**/37.48	**36.35**/35.64
v-a-t	47.34/34.50	47.38/34.97	38.22/24.20	45.73/31.40	**58.92**/59.32	53.38/59.32	45.14/45.69	**49.32**/49.74
a-t-v	**53.32**/39.25	**53.72**/40.42	43.26/30.10	47.65/32.88	57.97/59.63	**53.44**/59.76	51.22/52.49	45.00/45.13
a-v-t	51.54/37.36	50.45/36.47	46.24/32.73	**51.26**/34.85	54.32/55.89	50.00/56.98	55.27/55.99	40.41/39.64
Sarcasm Detection								
	MUStARD (2) Castro et al. (2019)				**MuSaG (2)** Scott et al. (2025)			
t-v-a	**78.16**/77.74	63.08/61.29	51.13/64.09	**56.32**/55.85	54.97/53.12	57.54/49.28	**53.80**/53.36	52.05/51.09
t-a-v	74.71/73.71	69.23/69.70	**70.67**/68.80	59.77/56.46	53.22/51.17	58.52/46.89	53.22/53.02	48.54/46.25
v-t-a	**70.11**/67.87	**61.54**/65.75	64.66/55.24	68.97/68.97	**50.29**/49.78	**51.47**/44.43	51.46/45.04	51.46/48.09
v-a-t	77.01/77.01	67.69/60.38	63.16/69.94	**75.86**/75.22	53.80/53.02	**60.13**/55.21	50.29/49.04	**42.11**/50.28
a-t-v	71.26/69.66	**75.38**/74.19	**47.37**/61.96	**56.32**/53.04	**55.56**/54.58	58.05/50.79	**49.12**/48.23	**52.63**/49.52
a-v-t	72.41/71.03	66.15/67.65	58.65/60.43	64.37/64.35	54.97/54.95	58.52/50.65	51.46/50.88	46.20/48.56
Humor Detection								
	UR-FUNNY (2) Hasan et al. (2019)				**M2H2 (2)** Chauhan et al. (2021)			
t-v-a	**58.65**/58.62	60.96/60.59	59.66/59.64	60.56/60.27	57.71/55.61	**61.71**/54.81	**61.38**/57.48	**63.90**/43.49
t-a-v	56.54/55.49	**53.92**/50.41	59.36/59.35	59.66/58.65	**58.36**/56.22	53.39/53.87	60.63/58.27	66.32/47.95
v-t-a	55.33/51.56	55.33/48.50	58.65/58.15	**57.95**/57.07	**55.02**/54.21	55.16/55.53	59.44/56.14	**63.90**/45.58
v-a-t	56.24/52.19	55.94/49.54	65.29/65.10	60.46/59.72	57.07/56.23	52.87/53.60	57.50/55.44	**66.49**/53.86
a-t-v	55.84/54.83	54.43/47.23	**52.92**/52.44	**60.97**/60.06	57.50/56.23	**52.49**/53.21	55.34/54.84	65.11/48.48
a-v-t	**51.81**/41.53	**62.47**/61.74	**66.50**/66.50	58.35/55.14	56.42/55.23	53.25/53.32	**52.75**/53.35	65.98/41.62
Hate Speech Detection								
	HateMM (2) Das et al. (2023)				**MultiHateClip (English) (2)** Wang et al. (2024)			
t-v-a	85.33/85.29	76.00/74.70	80.27/79.43	81.60/79.64	69.13/60.26	**62.95**/50.91	68.59/49.14	67.05/58.71
t-a-v	**87.33**/86.71	76.00/72.43	81.07/80.36	**85.33**/84.87	69.39/64.06	**65.24**/54.70	66.04/58.45	68.79/61.56
v-t-a	**82.67**/81.61	**72.00**/69.86	79.47/78.78	77.73/74.74	67.65/57.65	**65.24**/54.90	68.99/53.03	**66.71**/55.17
v-a-t	84.00/82.90	**77.33**/75.95	**83.20**/82.98	81.73/80.62	66.17/52.37	64.43/55.55	**64.97**/52.28	**71.41**/64.42
a-t-v	**82.67**/81.84	**72.00**/69.36	**76.00**/71.67	**77.33**/75.11	**69.80**/58.38	64.43/55.55	68.05/48.13	67.05/59.69
a-v-t	85.33/84.44	**77.33**/76.92	82.93/82.28	83.07/81.36	**64.16**/51.82	64.70/53.88	**71.01**/53.26	71.34/64.03
Deception Detection								
	RLTD (2) Pérez-Rosas et al. (2015)				**MU3D (2)** Lloyd et al. (2019)			
t-v-a	61.54/59.56	60.00/58.71	56.92/55.45	60.00/57.93	**56.87**/51.91	**43.75**/42.65	**53.12**/46.84	**50.62**/48.58
t-a-v	**66.15**/64.06	65.39/63.67	**58.46**/51.05	**63.08**/58.44	52.50/46.99	49.38/47.41	**61.54**/61.54	50.94/50.11
v-t-a	**55.39**/52.69	50.00/48.63	**52.31**/46.72	**52.31**/49.51	52.50/49.21	**43.75**/42.41	53.85/45.83	53.12/51.43
v-a-t	56.92/55.35	**46.15**/39.79	53.85/46.87	53.85/51.22	50.62/36.04	46.25/45.59	**61.54**/60.61	53.44/51.35
a-t-v	63.08/61.99	**66.15**/65.24	**58.46**/54.86	62.31/60.97	**49.38**/42.12	**51.88**/50.67	**61.54**/57.52	53.75/52.75
a-v-t	60.00/55.07	57.69/53.22	55.38/50.34	61.54/60.85	50.62/43.24	48.13/46.88	56.25/49.09	**54.06**/52.00

Table 2: Numerical accuracy (%) under different modality injection orders across four task families and four backbone models (P = Phi-3.5, Q = Qwen2.5, L = LLaVA-1.5, G = Gemma-3). For each order category, accuracy is averaged over all corresponding permutations and datasets. Color indicates relative performance within the same modality perspective: **high**, **mid**, **low**.

Order	Sentiment				Emotion				Sarcasm				Humor			
	P	Q	L	G	P	Q	L	G	P	Q	L	G	P	Q	L	G
t-first	46.14	43.57	46.73	47.85	41.78	42.15	49.50	42.48	65.27	62.09	57.21	54.17	57.82	57.50	60.26	62.61
t-mid	51.25	49.24	49.65	51.61	52.94	51.10	41.06	43.08	61.81	61.61	53.15	57.35	55.92	54.35	56.59	61.98
t-last	54.45	50.28	53.21	50.82	53.03	50.30	46.22	46.68	64.55	63.12	55.89	57.14	55.39	56.13	60.51	62.82
v-first	51.99	50.07	51.09	51.34	51.68	49.50	38.28	43.68	62.80	60.21	57.39	59.60	55.92	54.83	60.22	62.20
v-mid	51.64	47.49	50.51	50.27	46.93	45.44	52.32	43.32	65.13	61.32	53.76	54.74	56.15	59.60	60.07	62.20
v-last	48.22	45.53	48.00	48.67	49.14	48.62	46.18	45.24	63.69	65.30	55.10	54.32	57.06	53.56	57.06	63.02
a-first	53.72	49.45	51.77	51.08	54.29	51.90	49.00	46.08	63.55	64.53	51.65	54.88	55.39	55.66	56.88	62.60
a-mid	48.95	46.36	49.44	48.40	47.88	47.02	43.40	45.84	64.69	63.89	59.34	56.57	57.05	54.03	60.70	63.23
a-last	49.18	47.28	48.38	50.79	45.58	44.63	44.38	40.32	63.38	58.41	55.26	57.20	56.68	58.29	59.78	61.58

At the dataset level, IEMOCAP and MELD exhibit broadly consistent ordering preferences despite differing label spaces and recording conditions. For IEMOCAP, audio-first orders (*a-t-v* and *a-v-t*) are the most reliable configuration across models, whereas for MELD the advantage is somewhat more distributed between audio-first and vision-delayed configurations. The LLaVA-1.5 exception is particularly notable: its vision-middle preference stands in contrast to the other three models and may reflect model-specific interactions.



(a) MELD (Sentiment, 3-class) (b) MELD (Emotion, 7-class) (c) MUStARD (Sarcasm, 2-class) (d) UR-FUNNY (Humor, 2-class)

Figure 2: Effect of modality injection order on accuracy across four tasks and four datasets.

Sarcasm Detection. For MUStARD Castro et al. (2019) (Figure 2c), the *t-v-a*, *v-t-a*, and *a-t-v* orders generally under-perform. For MuSaG Scott et al. (2025), text-first and audio-first orders tend to outperform vision-first configurations. Across both datasets, text-middle orders (*v-t-a* and *a-t-v*) consistently achieve the lowest relative performance, suggesting a configuration advantage for early textual scheduling in sarcasm detection.

This contrasts sharply with sentiment and emotion analysis, where text-first orders are consistently disfavored. In sarcasm detection, early linguistic grounding appears to benefit the model by establishing the semantic context against which incongruent acoustic or visual cues can be interpreted. The consistent disadvantage of text-middle orders holds across both English (MUStARD) and German (MuSaG), suggesting the observed pattern is not specific to a single language. Vision-first configurations also tend to underperform, which aligns with the interpretation that visual cues alone, without prior linguistic context, provide insufficient grounding for detecting verbal irony.

Humor Detection. The two humor datasets reveal contrasting dataset-level preferences. UR-FUNNY Hasan et al. (2019) exhibits a clear tendency toward audio-first configurations: the *a-v-t* order achieves the strongest results for Qwen2.5 and LLaVA-1.5, while *a-t-v* leads for Gemma-3. Only Phi-3.5 deviates, favoring *t-v-a*. Vision-first orders consistently underperform on UR-FUNNY, suggesting that facial expressions and gestures provide supplementary rather than foundational signals for humor detection in monologic presentation settings. This audio prioritization aligns with the

dataset's origin in spoken TED-style presentations, where prosodic cues serve as primary markers of humorous intent.

By contrast, M2H2 Chauhan et al. (2021) favors text-first configurations: *t-v-a* achieves the highest accuracy for Qwen2.5 and LLaVA-1.5, while *t-a-v* leads for Phi-3.5. This divergence reflects the nature of scripted multiparty dialogues in M2H2, where humor comprehension depends more heavily on semantic content and conversational context. Across both datasets, text-middle orders (*v-t-a* and *a-t-v*) consistently yield the lowest performance in three of four models (Figure 2d) and the overall preference for non-middle textual scheduling aligns closely with the pattern observed in sarcasm detection, consistent with both tasks benefiting from either early or late semantic grounding.

Hate Speech Detection. For HateMM Das et al. (2023), text-middle configurations consistently under-perform orders in which text is injected first or last across all backbone models. For MultiHateClip Wang et al. (2024), whose noisy ASR-derived transcripts diminish the textual signal, this pattern is less pronounced and performance is more uniform across orders. Jointly, text-middle orders exhibit the weakest overall performance in three of four models, while vision-middle and audio-middle orders achieve stronger results.

The divergence between HateMM and MultiHateClip is informative: when the textual modality carries a strong, clean discriminative signal (HateMM), modality order effects are sharper and text-middle placement is clearly penalized. When transcript quality is degraded (MultiHateClip), the ordering advantage becomes diluted, as no single modality dominates and order effects are attenuated by noise. This suggests that the strength of observed order preferences may be partially mediated by the reliability of the individual modality signals, a pattern consistent with related observations in sarcasm and humor detection where one modality tends to dominate the relevant task signal.

Deception Detection. For RLTD Pérez-Rosas et al. (2015), vision-first orders consistently yield the lowest performance, whereas vision-last orders lead to the strongest results. Performance on MU3D Lloyd et al. (2019) is generally lower and appears closer to random variation. Deception detection often involves subtle cues and ambiguous samples, where even human annotators may find it difficult to distinguish.

Overall, we observe a consistent empirical pattern in which vision-last configurations outperform vision-middle and vision-middle outperform vision-first orders across models. Under the present controlled setting, this suggests a configuration advantage for deprioritizing visual input in deception detection, where linguistic content appears to provide stronger guidance during earlier decoding stages. Early textual injection shows a consistent downstream preference similar to the pattern observed in hate speech detection, though we note that causal interpretation of these trends remains outside the present scope.

5.2 All-at-Input Baseline Analysis

Table 3 summarizes the results across six tasks and twelve datasets. Contrary to the hypothesis that deeper processing consistently improves multimodal integration, the best sequential order with highest accuracy almost systematically achieves higher performance than the All-at-Input baseline.

This result suggests that blindly increasing the processing depth of all modalities does not necessarily lead to better multimodal representations. Instead, introducing modalities at carefully selected layers can help the model better absorb informative signals while avoiding interference from less relevant modalities.

When injected early, less informative modalities may dominate intermediate representations and introduce noise that propagates through subsequent layers. Sequential injection mitigates this issue by allowing the model to first establish a stable representation from more informative modalities before incorporating additional signals. This observation provides practical implications that MOCA-MLLM can serve as a lightweight, architecture-consistent mechanism for modality scheduling during downstream fine-tuning, with potential token-efficiency benefits discussed qualitatively.

Table 3: Performance comparison between the All-at-Input baseline and the best sequential modality injection configuration. Each value reports classification accuracy (%) and **Bold** denotes the higher accuracy.

Model	Phi	Qwen	LLaVA	Gemma	Phi	Qwen	LLaVA	Gemma
Sentiment Analysis								
	MELD				**CMU-MOSI**			
Baseline	46.03	43.27	41.58	44.36	31.63	35.42	53.50	**57.14**
Sequence	**58.64**	**53.50**	**53.65**	**53.14**	**54.08**	**48.94**	**55.10**	54.24
Emotion Recognition								
	MELD				**IEMOCAP**			
Baseline	43.65	39.63	35.74	41.12	51.35	39.32	48.11	36.35
Sequence	**53.32**	**53.72**	**48.32**	**51.26**	**58.92**	**53.44**	**59.46**	**49.32**
Sarcasm Detection								
	MUStARD				**MuSaG**			
Baseline	58.62	47.06	37.94	41.92	42.43	44.87	41.43	39.57
Sequence	**78.16**	**75.38**	**70.67**	**75.86**	**55.56**	**60.13**	**53.80**	**52.63**
Humor Detection								
	UR-FUNNY				**M2H2**			
Baseline	52.31	**62.88**	49.60	49.30	58.04	55.45	55.88	58.36
Sequence	**58.65**	62.47	**66.50**	**60.97**	**58.36**	**61.71**	**61.38**	**66.49**
Hate Speech Detection								
	HateMM				**MultiHateClip(English)**			
Baseline	82.13	63.73	82.67	76.27	**70.60**	60.27	69.13	66.85
Sequence	**87.33**	**77.33**	**83.20**	**85.33**	69.80	**65.24**	**71.01**	**71.41**
Deception Detection								
	RLTD				**MU3D**			
Baseline	52.31	**69.33**	47.69	55.38	53.12	50.00	54.37	50.00
Sequence	**66.15**	66.15	**58.46**	**63.08**	**56.87**	**51.88**	**61.54**	**54.06**

6 Conclusion & Future Work

We conduct the first large-scale, controlled study of modality injection order during MLLM fine-tuning, enabled by our proposed unified order-aware framework *MOCA-MLLM*. Across diverse tasks and datasets, we observe consistent and task-dependent empirical patterns: affective recognition tasks such as sentiment and emotion show a clear preference for early visual or acoustic scheduling, whereas semantic-grounded tasks including sarcasm, humor, hate speech and deception detection are consistently associated with better performance under early textual injection, with text-middle configurations emerging as the most suboptimal setting.

The proposed MOCA-MLLM framework enables isolating order effects from encoder bias and fusion mechanisms in a lightweight, architecture-consistent manner, making these observations reproducible across backbone models and datasets. Comparisons with the *All-at-Input* baseline empirically demonstrate that increasing the processing depth of all modalities does not necessarily lead to improved performance. Strategically scheduling when different modalities are introduced allows the model to better absorb informative signals while reducing interference from less relevant inputs.

Our findings highlight modality injection order as a critical yet previously under-explored design dimension, offering practical guidance for future MLLM development and task-specific fine-tuning. We additionally identify several promising directions for future research, which stay outside the scope of the present work. These include: (1) mechanistic investigations into the internal cause behind these empirical patterns; (2) generalizability to generative or open-ended tasks; (3) quantitative analysis of the potential token-efficiency benefits; (4) broader multilingual benchmarks, building upon our current benchmark that already includes non-English datasets such as MuSaG (German) and M2H2 (Hindi).

7 Reproducibility Statement

Codes are available at https://anonymous.4open.science/r/MOCA-MLLM-1D89/. The complete MOCA-MLLM procedure is summarized in Algorithm 1. More details regarding models, datasets, and reproducibility are discussed in Appendix C, D, and G respectively.

8 Ethics Statement

The datasets applied in Hate Speech Detection task contain offensive contents considering the nature of the classification task. Since we are using open-source dataset, we do not perform additional actions to counter the possible offenses.

9 AI Use Disclosure

In this work, we used generative AI tools for: assist with translation. We have **not** used generative AI tools for: generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, propose or refine hypotheses, design or provide feedback on research methodology or experiments, interpret results. Additionally, we used generative AI tools for: edit a research paper to improve readability, identify relevant literature. We have reviewed all AI-assisted work. All AI-assisted content, including literature suggestions and drafted text, was manually reviewed and revised by the authors. All algorithmic designs, implementations, experiments, and scientific conclusions were developed and validated by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahmoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Xihui (Eric) Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Olatunji Ruwase, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. Technical Report MSR-TR-2024-12, Microsoft, August 2024. URL https://www.microsoft.com/en-us/research/publication/phi-3-technical-report-a-highly-capable-language-model-locally-on-your-phone/.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, 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 *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International*

Conference on Neural Information Processing Systems, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, December 2008.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4619–4629. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1455. URL https://doi.org/10.18653/v1/p19-1455.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), March 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL https://doi.org/10.1145/3641289.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Navonil Majumder, Amir Zadeh, Asif Ekbal, Pushpak Bhattacharyya, Louis-philippe Morency, and Soujanya Poria. M2h2: A multimodal multiparty hindi dataset for humor recognition in conversations. In *Proceedings of the 2021 international conference on multimodal interaction*, pp. 773–777, 2021.

Yuzhong Chen, Hongren Pu, and Yang Qu. An analysis of attention mechanisms and its variance in transformer. *Applied and Computational Engineering*, 47(1):164–176, 2024a. ISSN 2755-2721.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pp. 1014–1023, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

FFmpeg Developers. Ffmpeg. https://ffmpeg.org/, 2024. Software.

Matt Gardner, Yoav Artzi, Jonathan Berant, Ben Bogin, Sihao Chen, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings*, 2020. URL https://api.semanticscholar.org/CorpusID:222124366.

Bryan Guan, Tanya Roosta, Peyman Passban, and Mehdi Rezagholizadeh. The order effect: Investigating prompt sensitivity to input order in llms, 2025. URL https://arxiv.org/abs/2502.04134.

Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan Kankanhalli, and Alberto Del Bimbo. On modality bias recognition and reduction. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(3), February 2023. ISSN 1551-6857. doi: 10.1145/3565266. URL https://doi.org/10.1145/3565266.

Aditi Gupta and Yossi Gandelsman. Test-time training for modality order consistency in vision-language models, 2026. URL https://arxiv.org/abs/2607.20351.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2046–2056, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1211. URL https://www.aclweb.org/anthology/D19-1211.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, and Lizhuang Ma. Efficient multimodal large language models: a survey. *Visual Intelligence*, 3(1):27, December 2025.

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*, ICML'23. JMLR.org, 2023.

Weixin Liang, LILI YU, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Nu6N69i8SB.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić (eds.), *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10/.

E Paige Lloyd, Jason C Deska, Kurt Hugenberg, Allen R McConnell, Brandon T Humphrey, and Jonathan W Kunstman. Miami university deception detection database. *Behavior research methods*, 51(1):429–439, 2019.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. 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. 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2025. URL https://arxiv.org/abs/2402.06196.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441/.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pp. 59–66, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450339124. doi: 10.1145/2818346.2820758. URL https://doi.org/10.1145/2818346.2820758.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Ali Reza Sajun, Imran Zualkernan, and Donthi Sankalpa. A historical survey of advances in transformer architectures. *Applied Sciences*, 14(10), 2024. ISSN 2076-3417. doi: 10.3390/app14104316. URL https://www.mdpi.com/2076-3417/14/10/4316.

Aaron Scott, Maike Züfle, and Jan Niehues. MuSaG: A Multimodal German Sarcasm Dataset with Full-Modal Annotations, 2025. URL https://arxiv.org/abs/2510.24178.

Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. Order matters: Exploring order sensitivity in multimodal large language models, 2024. URL https://arxiv.org/abs/2410.16983.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian

Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Mohit Tomar, Abhisek Tiwari, Tulika Saha, and Sriparna Saha. Your tone speaks louder than your face! modality order infused multi-modal sarcasm detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 3926–3933, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612528. URL https://doi.org/10.1145/3581783.3612528.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 200–212. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/01b7575c38dac42f3cfb7d500438b875-Paper.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=2ezugTT9kU.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 7493–7502, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681521. URL https://doi.org/10.1145/3664647.3681521.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

Grant Wardle and Teo Sušnjak. Image first or text first? optimising the sequencing of modalities in large language model prompting and reasoning tasks. *Big Data and Cognitive Computing*, 9 (6), 2025. ISSN 2504-2289. doi: 10.3390/bdcc9060149. URL https://www.mdpi.com/2504-2289/9/6/149.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Wenli Yang, Lilian Some, Michael Bain, and Byeong Kang. A comprehensive survey on integrating large language models with knowledge-based methods. *Knowledge-Based Systems*, 318:113503, 2025. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2025.113503. URL https://www.sciencedirect.com/science/article/pii/S0950705125005490.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision), 2023. URL https://arxiv.org/abs/2309.17421.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 11 2024. ISSN 2095-5138. doi: 10.1093/nsr/nwae403. URL https://doi.org/10.1093/nsr/nwae403.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016. URL https://arxiv.org/abs/1606.06259.

Yifan Zhan, Rui Yang, Junxian You, Mengjie Huang, Weibo Liu, and Xiaohui Liu. A systematic literature review on incomplete multimodal learning: techniques and challenges. *Systems Science & Control Engineering*, 13(1):2467083, 2025. doi: 10.1080/21642583.2025.2467083. URL https://doi.org/10.1080/21642583.2025.2467083.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pp. 12697–12706. PMLR, 2021.

Haohan Zheng and Zhenguo Zhang. Modality bias in LVLMs: Analyzing and mitigating object hallucination via attention lens, 2025. URL https://arxiv.org/abs/2508.02419.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-Zero's "Aha Moment" in Visual Reasoning on a 2B Non-SFT Model, 2025. URL https://arxiv.org/abs/2503.05132.

A FREQUENTLY ASKED QUESTIONS

Q1: Does the validity of the study depend on the specific fusion strategy used in MOCA-MLLM?

Response: Our design is consistent with the shared structural principle of the four evaluated backbones rather than being detached from mainstream MLLMs. Across Microsoft Phi-3.5-Vision Abdin et al. (2024), Alibaba Qwen2.5-Omni Xu et al. (2025), Meta LLaVA-1.5 Liu et al. (2023) and Google Gemma-3 Team et al. (2025), multimodal inputs are first encoded by modality-specific encoders and then projected into the language model hidden space before entering the main Transformer stack. None of these publicly described architectures relies on a dedicated standalone fusion stage placed entirely before the decoder blocks.

Our unified preprocessing pipeline follows this common interface: visual/audio representations are standardized into the same hidden space as textual tokens, while the original decoder layers are fully preserved. Therefore, regardless of whether a backbone internally uses self-attention interaction, gating, token mixing, learned resampling or architecture-specific routing inside the decoder, these mechanisms remain unchanged in our experiments. In this sense, our study isolates *when* modalities are introduced rather than redesigning *how* the pretrained decoder fuses information.

Furthermore, many practical MLLMs such as Qwen and LLaVA commonly reserve placeholders such as `[IMG]`, into which projected visual tokens are inserted into the textual sequence. The resulting multimodal sequence is then processed by standard Transformer attention layers. From an architectural perspective, this remains fundamentally a token insertion/concatenation followed by internal decoder fusion mechanism. Hence, our framework should be viewed as a controlled and architecture-consistent abstraction, rather than a departure from mainstream multimodal design.

Q2: Are the observed effects due to modality order or simply differences in processing depth?

Response: In this experimental setting, modality order is implemented through sequential injection at different decoder depths rather than token-level permutations at a single layer. Consequently, different modality orders naturally correspond to different depth allocations, and order and depth are partially entangled by design.

Nevertheless, depth alone cannot fully account for the observed patterns. The depth benefit applies symmetrically to whichever modality is injected first: in sentiment analysis, for instance, any first-injected modality receives the same number of subsequent decoder layers. Yet we consistently observe that vision-first and audio-first configurations outperform text-first configurations under this equal-depth condition. This suggests that modality identity and task-specific interaction patterns also contribute to the observed performance differences beyond depth alone. We report this strictly as an empirical observation within the present framework, without claiming a mechanistic causal explanation.

Importantly, the current MOCA-MLLM design can itself be interpreted as a structured ablation relative to the All-at-Input baseline. Different sequential configurations effectively delay specific modalities to later decoder stages while preserving the original decoder architecture, allowing us to directly evaluate how staged modality exposure influences downstream adaptation. In this sense, the present experimental setup already functions as a meaningful depth-aware ablation study and produces consistent empirical conclusions across models and tasks.

Previous work has also investigated modality order through sequential modality infusion at different encoder layers rather than token-level prompt ordering, demonstrating that such architectural scheduling is itself a meaningful research problemTomar et al. (2023). Additional ablations involving alternative spacing strategies or finer-grained injection intervals could certainly provide further insight and serve as useful robustness checks. However, such extensions are complementary rather than necessary for establishing the core empirical findings reported in this work.

Q3: Why is the analysis of the underlying reasons behind the observed modality order patterns limited?

Response: The primary objective of this work is to establish robust empirical trends under controlled experimental settings rather than to provide mechanistic explanations of multimodal integration. This is an intentional scope decision. Post-hoc explanations are offered as working hypotheses to

aid interpretation, but we avoid over-interpreting the evidence beyond what is directly supported by the experimental results.

We view the findings of this work as establishing reliable empirical regularities that provide a foundation for future mechanistic investigations into the architectural underpinnings of modality scheduling effects. Such investigations are outside the present scope but represent a natural and valuable extension.

Q4: To what extent do the findings generalize beyond the current MOCA-MLLM framework?

Response: We acknowledge that in order to construct a controlled large-scale empirical study, several standardized preprocessing choices were intentionally adopted. These include midpoint-frame visual selection, ASR-generated textual transcripts, pre-extracted audio embeddings and feature compression/projection pipelines. We fully recognize that such preprocessing decisions can influence absolute downstream performance and may also impose certain limitations on the generalization scope of the present findings. Different input representations, temporal sampling strategies or feature extraction pipelines could potentially lead to different quantitative results under alternative experimental settings.

However, we respectfully emphasize that when all modality-order configurations share the same preprocessing pipeline, the consistent modality-order patterns observed in our experiments remain compelling and empirically effective. Establishing such robust relative performance trends under controlled settings forms the primary objective and central contribution of this work.

Accordingly, we intentionally avoid making strong statements such as one modality being intrinsically more important than another. Instead, we frame our conclusions more carefully as empirical observations that certain modality injection sequences exhibit advantages over others for specific downstream tasks under consistent experimental conditions.

Q5: What is the practical value of the proposed MOCA-MLLM framework?

Response: We fully agree that for large-scale pretraining of foundation MLLMs, early fusion remains the dominant paradigm. However, this is not the setting our paper is intended to address. Our contribution focuses on downstream adaptation under limited fine-tuning resources. Compared with the conventional **All-at-Input** baseline, our study demonstrates that selectively injecting different modalities at different depths consistently leads to better downstream performance. Therefore, given a specific downstream task, practitioners can leverage our empirical observations to determine where each modality should be introduced, achieving stronger performance than injecting all modalities from the input layer while requiring only lightweight downstream fine-tuning rather than expensive multimodal pretraining. We therefore believe that the practical value of our findings within the intended scope of the paper is both clear and well supported by the experimental evidence.

Our primary finding is not merely that *order matters*, but that the preferred modality injection order exhibits consistent **task-dependent** patterns across different MLLM backbones. Based on these observations, we further identify two broad categories of downstream tasks: affective recognition and semantic-grounded tasks. Once the downstream task is known, practitioners can leverage these empirical observations to select a more appropriate injection schedule instead of adopting the conventional **All-at-Input** strategy by default. While our work does not aim to predict the optimal permutation for every individual dataset, we believe that establishing these consistent task-level patterns already provides meaningful and practical guidance.

B Pseudo Code

The pseudo code for the overall workflow is illustrated in Algorithm 1.

C Model Detail

We evaluate four SOTA MLLMs as shown in Table 4, serving as the backbone for our progressive fusion experiments.

Table 4: Summary of backbone MLLMs used in our experiments. For each model, we report the parameter scale, total number of transformer layers, modality injection points, and the layers unfrozen during fine-tuning.

Model(s)	Parameters	#Layers	Injection Points	Unfrozen Layers
Phi-3.5-Vision-Instruct	4.2B	32	L_1=11, L_2=22	Layers: 0–7, 11–18, 22–29
Qwen2.5-Omni-7B	7B	36	L_1=12, L_2=24	Layers: 0–3, 12–15, 24–27
LLaVA-1.5-7B	7B	32	L_1=11, L_2=22	Layers: 0–3, 11–14, 22–25
Gemma-3-4B-IT	4B	34	L_1=12, L_2=23	Layers: 0–5, 12–17, 23–28

D Dataset-Specific Strategies

D.1 Preprocessing Details

We perform a standardized preprocessing procedure to transform raw datasets into aligned modalities prior to encoder processing:

• *Audio & Visual Extraction*: We utilize `FFmpeg` to decompose raw video data FFmpeg Developers (2024). The midpoint frame is extracted to serve as the representative static image, while audio tracks are extracted and resampled to WAV format.

• *Text Transcription*: For samples lacking ground-truth textual transcriptions, we employ *WhisperX large-v2* to generate high-quality text transcripts from the audio input Bain et al. (2023).

• *Data Partitioning*: For datasets without pre-defined splits, we apply a stratified random split of *8:1:1* for training, validation, and testing.

D.2 Dataset Size

We adopt specific strategies to different datasets:

• *Standard Datasets*: We employ early stopping with a patience of 3 epochs, monitoring validation accuracy to prevent overfitting.

• *Small Datasets* (< 2000 samples): Acknowledging that validation sets in low-resource environments can be unreliable, we disable early stopping. Instead, we train for a fixed number of epochs determined empirically to reach convergence.

D.3 Label Mapping

CMU-MOSI Zadeh et al. (2016) provides continuous sentiment intensity annotations in the range of $[-3, 3]$, reflecting fine-grained emotional polarity. Following common practice to reduce label ambiguity near zero, we map the continuous scores into three sentiment classes using thresholds at ± 0.5, where scores greater than 0.5 are labeled as positive, scores lower than -0.5 as negative and the remaining samples as neutral.

IEMOCAP Busso et al. (2008) originally contains 11 emotion categories, yet several of them are extremely underrepresented. For stable optimization and unbiased decision boundaries, we address this severe class imbalance by merging and leaving out classes appropriately. Concretely, we combine *anger* (ang) and *disgust* (dis) as ang; *fear* (fea) and *frustration* (fru) as fru; *happiness* (hap), *excitement* (exc) and *surprise* (sur) as hap. We keep *sadness* (sad) and *neutral* (neu), yet we leave out *other* (oth) and *xxx* (xxx) for their lack of semantic representation. Eventually, we apply the 5-class dataset for downstream classification.

D.4 Pre-Extracted Embedding

We attempted to retrieve raw-source data for all benchmarks. However, datasets such as CMU-MOSI, whose original video samples are based on externally hosted YouTube content, often encounters substantial practical issues, including regional restrictions, unavailable links, removed videos,

Algorithm 1: Order-wise Modality Injection for Multimodal LLM Adaptation

Input : Multimodal inputs $\{x_m\}_{m\in\{t,v,a\}}$; pretrained encoders $\{E_m\}$; decoder f with L layers; modality order $\pi = (m_1, m_2, m_3)$

Output: Prediction logits $\hat{y}$

Freeze all pretrained encoders $\{E_m\}$ and all decoder parameters; `// Projection layers` $\{W_m, b_m\}$ `are unfrozen for training`

for *each modality* $m \in \{t, v, a\}$ **do**

  $H_m \leftarrow E_m(x_m)$; `// Extract embeddings`

  $\tilde{H}_m \leftarrow W_m H_m + b_m$; `// Project to` d_{LLM}

Set injection layers: $L_0 = 0$, $L_1 = \lceil L/3 \rceil$, $L_2 = \lceil 2L/3 \rceil$;

Initialize: $X^{(L_0)} \leftarrow \tilde{H}_{m_1}$;

for $\ell = L_0$ **to** $L_1 - 1$ **do**

  $X^{(\ell+1)} \leftarrow f_\ell(X^{(\ell)})$;

$X^{(L_1)} \leftarrow [X^{(L_1)}; \tilde{H}_{m_2}]$; `// Inject` m_2

for $\ell = L_1$ **to** $L_2 - 1$ **do**

  $X^{(\ell+1)} \leftarrow f_\ell(X^{(\ell)})$;

$X^{(L_2)} \leftarrow [X^{(L_2)}; \tilde{H}_{m_3}]$; `// Inject` m_3

for $\ell = L_2$ **to** $L - 1$ **do**

  $X^{(\ell+1)} \leftarrow f_\ell(X^{(\ell)})$;

Unfreeze equal numbers of decoder layers after each injection block, projection layers $\{W_m, b_m\}$, and the classification head;

$\hat{y} \leftarrow \text{Classifier}(X^{(L)})$;

return $\hat{y}$

and other platform-dependent inconsistencies. These factors can significantly degrade dataset completeness and reproducibility.

After carefully considering this tradeoff, we adopted a widely used and pragmatic practice in multimodal benchmark research: for datasets that provide official pre-extracted features, specifically CMU-MOSI, MUStARD Castro et al. (2019), UR-FUNNY Hasan et al. (2019) and M2H2 Chauhan et al. (2021), we utilize the provided embeddings directly; however, modality-specific projection layers are still applied prior to training to map all inputs into a unified LLM-compatible representation space. This is standard benchmark practice motivated by reproducibility and source availability constraints. Critically, this does not affect the validity of within-dataset comparisons across injection orders, which are the primary focus of our analysis, since all order configurations use the same pre-extracted features within each dataset.

E Statistical Analysis

Table 5 reports mean accuracy with standard deviation over 20 independent runs for all evaluated configurations, providing a basis for assessing the reliability of the observed trends across modality injection orders.

To further verify that the reported modality order preferences are statistically meaningful, we conduct paired two-sided t-tests comparing the accuracy distributions (20 runs each) of representative modality-order configurations for each task and backbone model. Concretely, we compare audio-first vs. text-first for sentiment analysis and emotion recognition; text-first vs. text-middle for sarcasm, humor, and hate speech detection; vision-last vs. vision-first for deception detection; and the best sequential injection configuration against the All-at-Input baseline.

Table 6 reports the corresponding t-statistics, with significance levels $^{***}(p < 0.001)$, $^{**}(p < 0.01)$, and $^{*}(p < 0.05)$. The large majority of comparisons yield $|t| > 3$ at $p < 0.001$, confirming that the modality-order preferences identified in the main analysis are robust and consistent across repeated runs.

The LLaVA-1.5 emotion recognition comparison does not reach significance, reflecting the near-identical performance of audio-first and text-first configurations for that model. The Qwen2.5 sarcasm comparison is similarly non-significant, as text-first and text-middle configurations yield closely matched accuracy under that setting. The Gemma-3 sarcasm result is significant in the reverse direction ($t = -6.29^{***}$), confirming the text-middle advantage for that model documented in the task analysis. The Gemma-3 humor comparison does not reach significance due to a small mean difference between configurations. The LLaVA-1.5 hate speech comparison is likewise non-significant, consistent with the marginal gap reported between text-first and text-middle for that dataset. These exceptions are fully consistent with the per-model, per-task discussion and do not undermine the overall empirical trends.

F More Results and Visualizations

The complete results of 4 MLLMs, 6 downstream tasks, 12 multimodal datasets and 6 modality sequences are displayed in Tables 7 and 8, with evaluation criteria across accuracy, macro-precision, macro-recall, and macro-f1-score. Figures 3 and 4 visualize the effect of modality injection order on accuracy, showing order-wise performance trends for all backbone models.

Figures 5 to 7 visualize the effect of modality injection order on accuracy for each task group, illustrating modality-wise performance trends across text-, vision-, and audio-first configurations.

Additionally, Figures 8 and 9 present representative confusion matrices illustrating model behavior under selected modality injection orders.

Finally, Table 9 reports the complete comparisons between the All-at-Input baseline and the best sequential modality injection results across all evaluated tasks and backbone models. Table 5 reports mean accuracy with standard deviation over 20 independent runs for all configurations.

G Reproducibility Details

G.1 Computing Environment

All experiments were conducted on a server equipped with NVIDIA A800 GPU(s) with 80 GB memory. The core software environment consisted of Ubuntu 22.04.5 LTS, Python 3.12.3, CUDA 12.8, PyTorch 2.8.0 (CUDA 12.8 build), Hugging Face Transformers 4.50.0, and Accelerate 1.11.0.

G.2 Hyper-parameter Configuration

Hyper-parameters were tuned independently for each model–dataset pair. The following search ranges were considered where applicable:

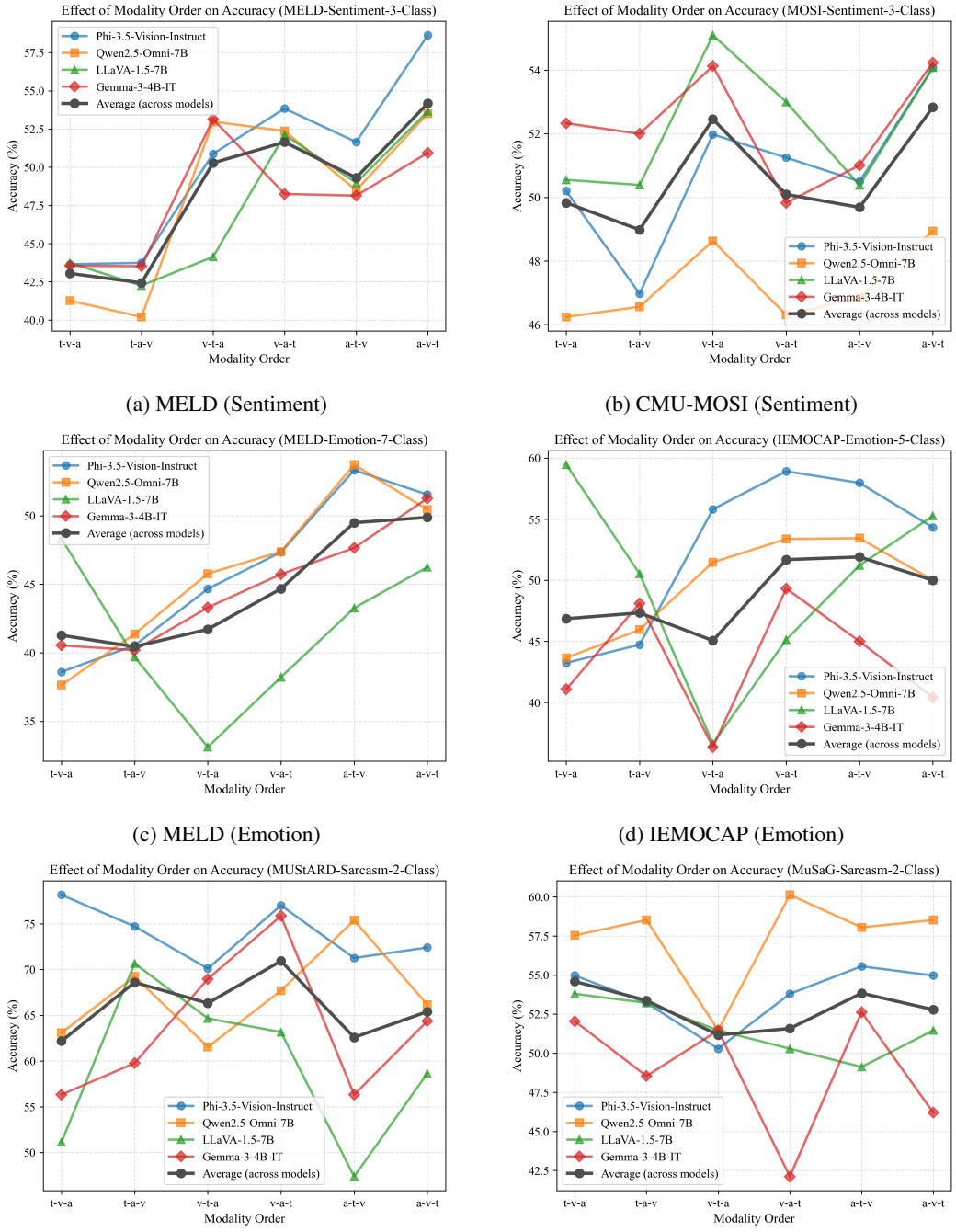
- Learning rate: $\{1 \times 10^{-4}, 5 \times 10^{-4}, 2 \times 10^{-5}, 5 \times 10^{-6}\}$
- Weight decay: $\{1 \times 10^{-4}, 5 \times 10^{-4}, 2 \times 10^{-5}, 5 \times 10^{-6}\}$
- Batch size: $4, 8, 16, 32$
- Number of epochs: 5, 10, or 10 with early stopping.

The final hyper-parameter configuration for each experiment was selected based on validation performance while following the official recommendations of the corresponding pretrained model whenever applicable.

H Artifact Usage

This work utilizes publicly available scientific artifacts, including benchmark multimodal datasets and pretrained models commonly adopted in multimodal machine learning research.

All utilized artifacts are appropriately cited in the main paper. These artifacts were used exclusively for academic research purposes and in a manner consistent with their intended research usage and public availability conditions.



(a) MELD (Sentiment)

(b) CMU-MOSI (Sentiment)

(c) MELD (Emotion)

(d) IEMOCAP (Emotion)

(e) MUStARD (Sarcasm)

(f) MuSaG (Sarcasm)

Figure 3: Effect of modality injection order on accuracy across sentiment analysis, emotion recognition, and sarcasm detection tasks. Each subplot shows accuracy trends under different modality orders for four multimodal backbone models.

Table 5: Performance of different modality injection orders across six tasks (number of classes indicated), twelve datasets, and four backbone models. Each cell reports mean accuracy±standard deviation (in %), computed over 20 independent runs. For each model and dataset, **Bold** (Plain) denotes the highest (second highest) mean accuracy, and **Bold** (Plain) denotes the lowest (second lowest) mean accuracy amongst all modality orders.

Order	Phi	Qwen	LLaVA	Gemma	Phi	Qwen	LLaVA	Gemma
Sentiment Analysis								
	MELD (3) Poria et al. (2019)				**CMU-MOSI (3)** Zadeh et al. (2016)			
t-v-a	43.65±3.03	41.26±2.31	43.74±2.70	43.55±2.61	50.20±2.13	**46.24±2.38**	50.55±1.03	52.33±2.04
t-a-v	**43.74±2.36**	**40.20±2.21**	**42.25±2.77**	**43.52±2.63**	**46.97±2.60**	46.56±2.45	50.39±0.89	52.00±1.26
v-t-a	50.88±2.52	52.98±1.93	44.13±2.88	**53.14±2.62**	51.98±2.27	48.63±1.89	**55.10±1.33**	54.13±2.43
v-a-t	53.84±3.02	52.37±2.49	52.12±2.54	48.25±2.23	51.25±1.15	46.30±1.60	53.00±0.85	**49.83±0.88**
a-t-v	51.65±2.15	48.49±2.30	48.98±2.10	48.14±1.99	50.50±2.53	46.85±2.08	**50.38±1.85**	51.01±1.87
a-v-t	**58.64±2.67**	**53.50±2.45**	**53.65±2.09**	50.94±2.24	**54.08±1.94**	**48.94±1.32**	54.08±2.15	**54.24±2.15**
Emotion Recognition								
	MELD (7)				**IEMOCAP (5)** Busso et al. (2008)			
t-v-a	**38.60±2.18**	**37.64±2.52**	**48.32±2.56**	40.54±2.49	**43.24±1.25**	**43.65±1.74**	**59.46±1.87**	41.08±1.85
t-a-v	40.54±2.86	41.37±2.48	39.69±3.10	**40.19±3.04**	44.73±0.95	45.95±2.59	50.54±2.38	48.11±0.83
v-t-a	44.66±2.78	45.77±2.08	**33.12±3.00**	43.30±2.04	55.81±1.86	51.47±1.14	**36.62±1.17**	**36.35±2.56**
v-a-t	47.34±2.80	47.38±2.20	38.22±2.41	45.73±2.42	**58.92±1.09**	53.38±2.07	45.14±2.57	**49.32±2.13**
a-t-v	**53.32±2.62**	**53.72±2.95**	43.26±2.42	47.65±2.11	57.97±1.39	**53.44±0.89**	51.22±2.15	45.00±1.24
a-v-t	51.54±2.30	50.45±2.73	46.24±2.77	**51.26±2.42**	54.32±1.43	50.00±1.91	55.27±1.90	40.41±2.18
Sarcasm Detection								
	MUStARD (2) Castro et al. (2019)				**MuSaG (2)** Scott et al. (2025)			
t-v-a	**78.16±1.29**	63.08±1.56	51.13±1.18	**56.32±1.92**	54.97±0.81	57.54±2.49	**53.80±1.28**	52.05±0.85
t-a-v	74.71±2.21	69.23±2.57	**70.67±1.87**	59.77±2.12	53.22±2.26	58.52±2.59	53.22±1.37	48.54±1.75
v-t-a	**70.11±1.37**	**61.54±1.69**	64.66±1.46	68.97±1.19	**50.29±1.57**	**51.47±0.98**	51.46±2.22	51.46±1.60
v-a-t	77.01±1.48	67.69±1.95	63.16±2.40	**75.86±0.96**	53.80±2.38	**60.13±1.35**	50.29±1.59	**42.11±0.95**
a-t-v	71.26±2.59	**75.38±2.42**	**47.37±1.65**	**56.32±1.12**	**55.56±1.35**	58.05±2.51	**49.12±2.02**	**52.63±1.65**
a-v-t	72.41±1.63	66.15±1.27	58.65±2.24	64.37±2.29	54.97±2.49	58.52±1.47	51.46±1.29	46.20±1.09
Humor Detection								
	UR-FUNNY (2) Hasan et al. (2019)				**M2H2 (2)** Chauhan et al. (2021)			
t-v-a	**58.65±1.64**	60.96±2.47	59.66±1.19	60.56±0.85	57.71±2.14	**61.71±1.09**	**61.38±2.38**	**63.90±2.16**
t-a-v	56.54±2.30	**53.92±1.76**	59.36±2.28	59.66±1.96	**58.36±1.81**	53.39±1.30	60.63±1.42	66.32±1.52
v-t-a	55.33±1.01	55.33±1.82	58.65±1.47	**57.95±1.92**	**55.02±1.92**	55.16±1.67	59.44±1.02	**63.90±1.13**
v-a-t	56.24±1.08	55.94±1.90	65.29±1.62	60.46±0.87	57.07±2.44	52.87±1.68	57.50±0.97	**66.49±1.45**
a-t-v	55.84±0.94	54.43±1.31	**52.92±0.96**	**60.97±2.05**	57.50±1.88	**52.49±2.09**	55.34±1.90	65.11±0.87
a-v-t	**51.81±1.70**	**62.47±1.08**	**66.50±2.30**	58.35±0.82	56.42±1.88	53.25±2.32	**52.75±1.03**	65.98±1.60
Hate Speech Detection								
	HateMM (2) Das et al. (2023)				**MultiHateClip (English) (2)** Wang et al. (2024)			
t-v-a	85.33±2.44	76.00±1.39	80.27±1.69	81.60±1.57	69.13±1.77	**62.95±1.78**	68.59±2.00	67.05±1.61
t-a-v	**87.33±1.01**	76.00±2.49	81.07±1.16	**85.33±1.22**	69.39±2.02	**65.24±2.35**	66.04±1.59	68.79±0.99
v-t-a	**82.67±1.27**	**72.00±1.16**	79.47±1.07	77.73±1.31	67.65±2.22	**65.24±1.91**	68.99±2.15	**66.71±2.59**
v-a-t	84.00±1.08	**77.33±1.46**	**83.20±1.77**	81.73±1.13	66.17±2.56	64.43±2.33	**64.97±2.10**	**71.41±1.24**
a-t-v	**82.67±1.57**	**72.00±1.44**	**76.00±1.26**	**77.33±2.59**	**69.80±2.31**	64.43±1.71	68.05±1.70	67.05±1.24
a-v-t	85.33±2.51	**77.33±2.45**	82.93±2.22	83.07±1.31	**64.16±0.81**	64.70±1.77	**71.01±1.71**	71.34±0.90
Deception Detection								
	RLTD (2) Pérez-Rosas et al. (2015)				**MU3D (2)** Lloyd et al. (2019)			
t-v-a	61.54±2.58	60.00±0.82	56.92±1.42	60.00±1.79	**56.87±2.19**	**43.75±2.02**	**53.12±1.66**	**50.62±1.33**
t-a-v	**66.15±1.25**	65.39±1.95	**58.46±1.50**	**63.08±1.07**	52.50±2.55	49.38±1.75	**61.54±1.85**	50.94±1.96
v-t-a	**55.39±1.22**	50.00±2.05	**52.31±2.38**	**52.31±2.12**	52.50±2.21	**43.75±1.18**	53.85±1.63	53.12±1.82
v-a-t	56.92±2.40	**46.15±1.98**	53.85±2.21	53.85±2.48	50.62±2.41	46.25±0.86	**61.54±0.97**	53.44±1.82
a-t-v	63.08±1.35	**66.15±1.63**	**58.46±1.72**	62.31±2.15	**49.38±1.19**	**51.88±2.06**	**61.54±1.82**	53.75±1.16
a-v-t	60.00±2.36	57.69±1.49	55.38±2.37	61.54±1.42	50.62±1.84	48.13±2.57	56.25±2.06	**54.06±0.99**

Table 6: Paired two-sided t-test statistics comparing representative modality-order configurations across backbone models (20 runs per configuration). Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Positive t-values indicate that the first-listed configuration achieves higher mean accuracy than the second-listed configuration.

Task	**Comparison**	**Phi-3.5**	**Qwen2.5**	**LLaVA-1.5**	**Gemma-3**
Sentiment	Audio-first vs. Text-first	11.42^{***}	8.86^{***}	7.59^{***}	4.87^{***}
Emotion	Audio-first vs. Text-first	14.14^{***}	11.02^{***}	0.57	4.07^{***}
Sarcasm	Text-first vs. Text-middle	5.21^{***}	0.95	8.02^{***}	-6.29^{***}
Humor	Text-first vs. Text-middle	3.75^{***}	6.23^{***}	7.25^{***}	1.25
Hate Speech	Text-first vs. Text-middle	4.74^{***}	3.68^{***}	1.94	7.86^{***}
Deception	Vision-last vs. Vision-first	7.29^{***}	14.74^{***}	8.57^{***}	8.63^{***}
All-at-Input	Best Seq. vs. All-at-Input	5.82^{***}	4.66^{***}	5.33^{***}	4.16^{***}

Table 7: Complete evaluation results for Sentiment Analysis, Emotion Recognition, and Sarcasm Detection (number of classes indicated), reporting accuracy (Acc), precision (P), recall (R), and F1-score (F1) (in %) under all modality injection orders across datasets and backbone models. For each model and dataset, **Bold** (Plain) denotes the highest (second highest) accuracy, and **Bold** (Plain) denotes the lowest (second lowest) accuracy amongst all modality orders.

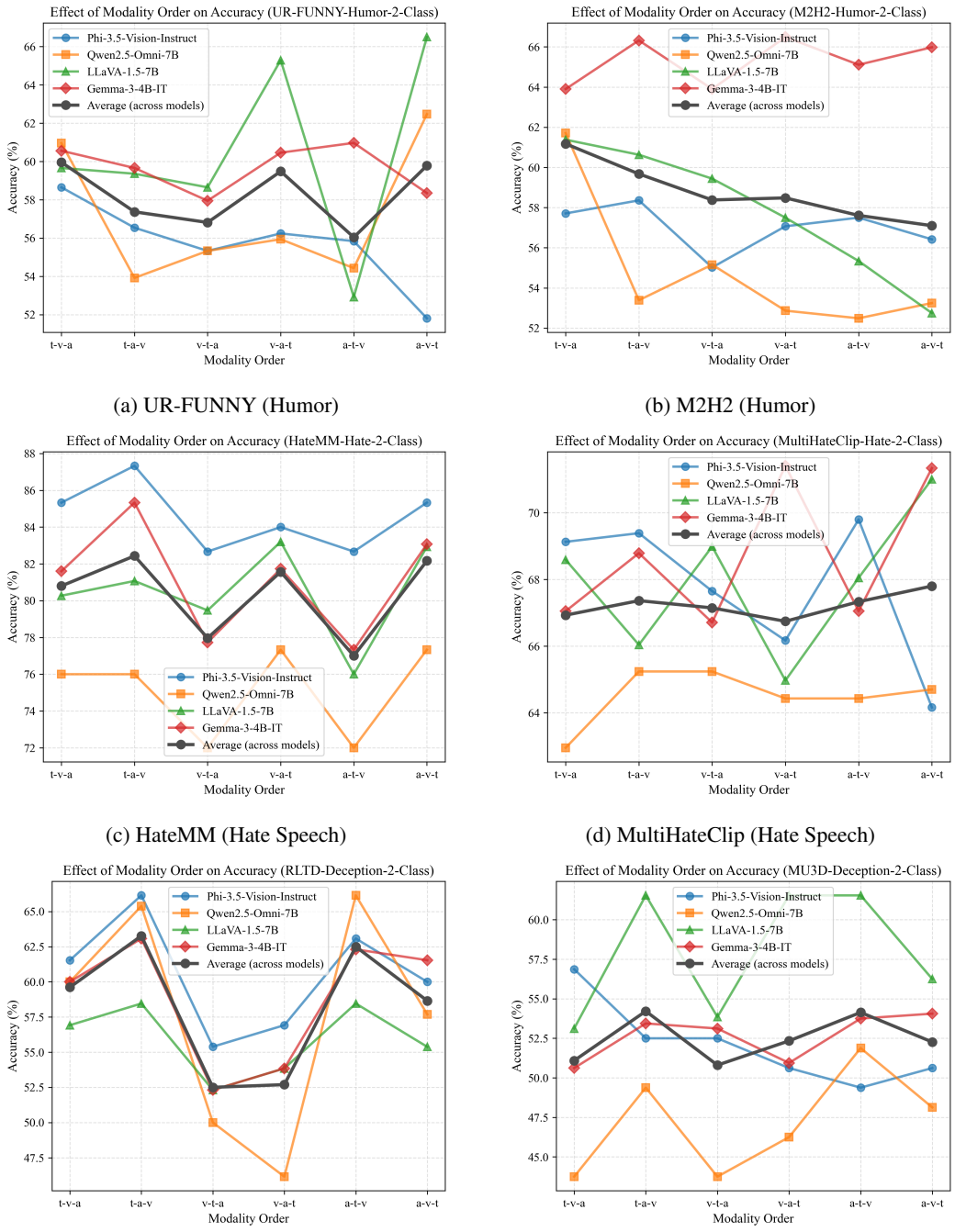
Order	Phi-3.5				Qwen2.5				LLaVA-1.5				Gemma-3			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Sentiment Analysis – MELD (3)																
t-v-a	43.65	28.31	30.79	29.52	41.26	25.23	27.42	26.30	43.74	28.65	33.10	28.69	43.55	25.04	26.23	25.63
t-a-v	**43.74**	30.14	28.32	29.21	**40.20**	25.44	26.85	26.12	**42.25**	28.44	25.13	27.84	**43.52**	30.32	22.65	26.12
v-t-a	50.88	35.28	36.49	35.87	52.98	37.33	35.10	36.19	44.13	23.34	25.12	26.03	**53.14**	40.33	35.14	37.61
v-a-t	53.84	38.37	40.34	39.34	52.37	36.83	38.81	37.80	52.12	29.56	33.73	33.04	48.25	31.55	32.32	31.93
a-t-v	51.65	35.04	36.13	35.58	48.49	33.90	34.12	34.01	48.98	33.23	32.66	33.60	48.14	33.52	33.66	33.59
a-v-t	**58.64**	37.39	38.25	37.82	**53.50**	37.55	35.23	36.36	**53.65**	37.44	37.98	38.04	50.94	35.11	36.24	35.67
Sentiment Analysis – CMU-MOSI (3)																
t-v-a	50.20	50.85	47.59	45.74	**46.24**	44.86	45.27	42.46	50.55	42.10	43.06	30.10	52.33	46.49	45.09	40.58
t-a-v	**46.97**	49.26	46.85	41.67	46.56	46.01	45.76	44.06	50.39	41.93	41.64	35.64	52.00	48.24	45.50	40.32
v-t-a	51.98	52.73	51.34	48.91	48.63	45.35	46.32	42.22	**55.10**	44.90	45.21	42.49	54.13	48.47	47.75	44.87
v-a-t	51.25	48.01	47.78	45.94	46.30	44.11	44.95	42.27	53.00	43.23	44.33	40.71	**49.83**	45.80	44.24	39.36
a-t-v	50.50	51.53	49.80	46.91	46.85	44.27	45.51	42.18	**50.38**	40.12	42.68	37.99	51.01	44.26	44.58	39.91
a-v-t	**54.08**	47.62	47.73	46.46	**48.94**	48.89	49.36	46.41	54.08	46.00	44.26	39.89	**54.24**	46.37	46.63	43.35
Emotion Recognition – MELD (7)																
t-v-a	**38.60**	25.66	25.51	25.58	**37.64**	23.14	22.96	23.05	**48.32**	32.44	35.71	34.03	40.54	25.10	24.98	25.04
t-a-v	40.54	28.72	24.81	26.68	41.37	28.19	28.33	28.26	39.69	24.60	25.65	25.12	**40.19**	26.19	26.24	26.21
v-t-a	44.66	30.92	32.93	31.91	45.77	31.23	33.68	32.43	**33.12**	19.02	18.85	18.93	43.30	28.86	29.32	29.09
v-a-t	47.34	34.33	34.67	34.50	47.38	36.69	33.36	34.97	38.22	23.62	24.79	24.20	45.73	31.34	31.47	31.40
a-t-v	**53.32**	38.44	40.08	39.25	**53.72**	39.54	41.32	40.42	43.26	29.87	30.33	30.10	47.65	32.55	33.21	32.88
a-v-t	51.54	28.42	26.34	27.36	50.45	26.76	26.18	26.47	46.24	22.14	23.34	22.73	**51.26**	24.15	25.56	24.85
Emotion Recognition – IEMOCAP (5)																
t-v-a	**43.24**	48.40	44.51	44.23	**43.65**	47.34	44.60	43.78	**59.46**	60.29	58.98	59.49	41.08	42.23	41.29	40.62
t-a-v	44.73	48.64	46.56	45.50	45.95	44.30	47.56	45.84	50.54	50.59	51.42	50.88	48.11	48.08	48.96	48.30
v-t-a	55.81	57.61	56.67	56.31	51.47	56.62	59.87	56.49	**36.62**	37.77	40.09	37.48	**36.35**	35.96	35.59	35.64
v-a-t	**58.92**	61.29	58.74	59.32	53.38	61.29	58.74	59.32	45.14	45.57	48.10	45.69	**49.32**	49.84	50.47	49.74
a-t-v	57.97	60.21	59.93	59.63	**53.44**	60.95	59.88	59.76	51.22	52.18	53.21	52.49	45.00	44.89	45.55	45.13
a-v-t	54.32	55.54	57.23	55.89	50.00	58.76	57.44	56.98	55.27	56.74	55.41	55.99	40.41	40.14	39.50	39.64
Sarcasm Detection – MUStARD (2)																
t-v-a	**78.16**	80.11	78.01	77.74	63.08	57.58	65.52	61.29	51.13	48.33	95.08	64.09	**56.32**	56.46	56.21	55.85
t-a-v	74.71	78.61	74.50	73.71	69.23	62.16	79.31	69.70	**70.67**	67.19	70.49	68.80	59.77	64.78	60.10	56.46
v-t-a	**70.11**	78.84	70.43	67.87	**61.54**	54.55	82.76	65.75	64.66	65.91	47.54	55.24	68.97	68.97	68.97	68.97
v-a-t	77.01	77.01	77.01	77.01	67.69	66.67	55.17	60.38	63.16	55.88	93.44	69.94	**75.86**	78.42	75.69	75.22
a-t-v	71.26	77.68	71.54	69.66	**75.38**	69.70	79.31	74.19	**47.37**	46.34	93.44	61.96	**56.32**	59.37	56.63	53.04
a-v-t	72.41	77.06	72.17	71.03	66.15	58.97	79.31	67.65	58.65	53.85	68.85	60.43	64.37	64.37	64.35	64.35
Sarcasm Detection – MuSaG (2)																
t-v-a	54.97	53.55	53.33	53.12	57.54	53.97	51.21	49.28	**53.80**	53.38	53.39	53.36	52.05	51.38	51.40	51.09
t-a-v	53.22	51.59	51.48	51.17	58.52	52.73	49.96	46.89	53.22	53.18	53.21	53.02	48.54	46.74	47.05	46.25
v-t-a	**50.29**	49.46	49.55	49.78	**51.47**	53.28	48.11	44.43	51.46	48.49	49.27	45.04	51.46	48.43	48.40	48.09
v-a-t	53.80	53.03	53.12	53.02	**60.13**	55.78	55.50	55.21	50.29	49.43	49.44	49.04	**42.11**	51.12	51.29	50.28
a-t-v	**55.56**	54.65	54.58	54.58	58.05	51.20	51.07	50.79	**49.12**	48.46	48.49	48.23	**52.63**	49.72	49.72	49.52
a-v-t	54.97	55.47	55.52	54.95	58.52	53.61	51.45	50.65	51.46	51.30	51.21	50.88	46.20	48.65	48.63	48.56

Table 8: Complete evaluation results for Humor Detection, Hate Speech Detection, and Deception Detection (number of classes indicated), reporting accuracy (Acc), precision (P), recall (R), and f1-score (F1) (in %) under all modality injection orders across datasets and backbone models. For each model and dataset, **Bold** (Plain) denotes the highest (second highest) accuracy, and **Bold** (Plain) denotes the lowest (second lowest) accuracy amongst all modality orders.

Order	Phi-3.5				Qwen2.5				LLaVA-1.5				Gemma-3			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Humor Detection – UR-FUNNY (2)																
t-v-a	**58.65**	58.74	58.65	58.62	60.96	61.58	60.97	60.59	59.66	59.74	59.70	59.64	60.56	60.72	60.45	60.27
t-a-v	56.54	57.51	56.54	55.49	**53.92**	56.06	53.92	50.41	59.36	59.40	59.38	59.35	59.66	60.37	59.45	58.65
v-t-a	55.33	57.12	55.33	51.56	55.33	62.67	55.33	48.50	58.65	59.32	58.82	58.15	**57.95**	58.37	57.76	57.07
v-a-t	56.24	60.17	56.24	52.19	55.94	60.87	55.94	49.54	65.29	65.81	65.41	65.10	60.46	61.61	60.67	59.72
a-t-v	55.84	56.16	55.84	54.83	54.43	61.08	54.43	47.23	**52.92**	52.91	52.91	52.44	**60.97**	62.43	61.19	60.06
a-v-t	**51.81**	58.32	51.81	41.53	**62.47**	63.79	62.47	61.74	**66.50**	66.53	66.52	66.50	58.35	62.51	58.74	55.14
Humor Detection – M2H2 (2)																
t-v-a	57.71	54.46	57.71	55.61	**61.71**	55.00	61.71	54.81	**61.38**	56.52	61.38	57.48	**63.90**	47.65	49.35	43.49
t-a-v	**58.36**	55.12	58.36	56.22	53.39	54.49	53.30	53.87	60.63	57.32	60.63	58.27	66.32	57.70	52.33	47.95
v-t-a	**55.02**	53.58	55.02	54.21	55.16	55.98	55.16	55.53	59.44	54.87	59.44	56.14	**63.90**	50.36	50.13	45.58
v-a-t	57.07	55.61	57.07	56.23	52.87	54.73	52.87	53.60	57.50	54.30	57.50	55.44	**66.49**	59.38	55.18	53.86
a-t-v	57.50	55.40	57.50	56.23	**52.49**	54.33	52.49	53.21	55.34	54.41	55.34	54.84	65.11	54.86	51.94	48.48
a-v-t	56.42	54.40	56.42	55.23	53.25	53.39	53.25	53.32	**52.75**	54.11	52.75	53.35	65.98	50.00	50.00	41.62
Hate Speech Detection – HateMM (2)																
t-v-a	85.33	85.68	86.41	85.29	76.00	76.22	74.27	74.70	80.27	80.65	79.35	79.43	81.60	85.30	80.00	79.64
t-a-v	**87.33**	86.71	86.80	86.71	76.00	82.47	72.27	72.43	81.07	84.18	81.41	80.36	**85.33**	85.79	84.89	84.87
v-t-a	**82.67**	84.00	80.89	81.61	**72.00**	72.53	69.59	69.86	79.47	79.46	79.14	78.78	77.73	82.58	75.62	74.74
v-a-t	84.00	86.03	82.05	82.90	**77.33**	78.00	75.44	75.95	**83.20**	83.29	83.35	82.98	81.73	83.51	80.83	80.62
a-t-v	**82.67**	83.22	81.29	81.84	**72.00**	73.28	69.19	69.36	**76.00**	81.19	72.51	71.67	**77.33**	82.56	76.36	75.11
a-v-t	85.33	87.00	83.61	84.44	**77.33**	76.84	77.03	76.92	82.93	83.13	81.92	82.28	83.07	86.23	81.40	81.36
Hate Speech Detection – MultiHateClip (English) (2)																
t-v-a	69.13	66.42	61.16	60.26	**62.95**	53.12	51.90	50.91	68.59	67.22	54.09	49.14	67.05	62.03	59.29	58.71
t-a-v	69.39	65.39	63.98	64.06	**65.24**	57.20	55.11	54.70	66.04	59.97	59.43	58.45	68.79	64.00	61.89	61.56
v-t-a	67.65	62.90	58.54	57.65	**65.24**	57.24	55.22	54.90	68.99	56.82	56.14	53.03	**66.71**	56.64	56.64	55.17
v-a-t	66.17	51.39	57.01	52.37	64.43	57.92	56.27	55.55	**64.97**	63.05	56.23	52.28	**71.41**	67.19	63.82	64.42
a-t-v	**69.80**	60.28	59.79	58.38	64.43	57.92	56.27	55.55	68.05	51.83	53.48	48.13	67.05	62.76	60.01	59.69
a-v-t	**64.16**	55.69	55.31	51.82	64.70	56.49	54.39	53.88	**71.01**	56.89	56.97	53.26	71.34	66.73	63.56	64.03
Deception Detection – RLTD (2)																
t-v-a	61.54	62.05	60.47	59.56	60.00	60.52	60.47	58.71	56.92	56.55	56.43	55.45	60.00	63.71	60.48	57.93
t-a-v	**66.15**	68.14	65.95	64.06	65.39	66.12	64.29	63.67	**58.46**	62.61	59.05	51.05	**63.08**	66.06	61.90	58.44
v-t-a	**55.39**	60.44	57.14	52.69	50.00	50.56	50.60	48.63	**52.31**	46.64	51.90	46.72	**52.31**	51.79	52.38	49.51
v-a-t	56.92	59.74	57.38	55.35	**46.15**	36.75	48.22	39.79	53.85	51.22	52.14	46.87	53.85	55.52	53.93	51.22
a-t-v	63.08	63.16	62.38	61.99	**66.15**	66.77	65.48	65.24	**58.46**	60.53	56.67	54.86	62.31	62.68	61.67	60.97
a-v-t	60.00	63.74	62.62	55.07	57.69	60.84	55.95	53.22	55.38	59.68	55.24	50.34	61.54	61.73	61.31	60.85
Deception Detection – MU3D (2)																
t-v-a	**56.87**	54.01	56.87	51.91	**43.75**	43.34	43.75	42.65	**53.12**	55.93	53.12	46.84	**50.62**	50.89	50.62	48.58
t-a-v	52.50	46.31	52.50	46.99	49.38	49.78	49.38	47.41	**61.54**	61.90	61.90	61.54	50.94	54.94	50.94	50.11
v-t-a	52.50	51.58	52.50	49.21	**43.75**	43.39	43.75	42.41	53.85	52.27	51.19	45.83	53.12	53.23	53.12	51.43
v-a-t	50.62	31.19	50.62	36.04	46.25	45.96	46.25	45.59	**61.54**	65.28	63.10	60.61	53.44	50.89	53.44	51.35
a-t-v	**49.38**	39.47	49.38	42.12	**51.88**	52.31	51.88	50.67	**61.54**	63.33	59.52	57.52	53.75	54.63	53.75	52.75
a-v-t	50.62	40.87	50.62	43.24	48.13	47.75	48.13	46.88	56.25	64.29	56.25	49.09	**54.06**	53.92	54.06	52.00

Table 9: Performance comparison between the All-at-Input baseline and the best sequential modality injection configuration, reporting accuracy (Acc), precision (P), recall (R), and f1-score (F1) (in %) under all modality injection orders across datasets and backbone models. For each model and dataset, **Bold** denotes the higher accuracy.

Order	Phi-3.5				Qwen2.5				LLaVA-1.5				Gemma-3			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Sentiment Analysis – MELD (3)																
Baseline	46.03	29.64	28.81	29.21	43.27	27.16	26.54	26.85	41.58	26.02	25.37	25.69	44.36	27.91	27.24	27.56
Sequence	**58.64**	37.39	38.25	37.82	**53.50**	37.55	35.23	36.36	**53.65**	37.44	37.98	38.04	**53.14**	40.33	35.14	37.61
Sentiment Analysis – CMU-MOSI (3)																
Baseline	31.63	33.66	33.52	29.69	35.42	29.23	33.07	24.87	53.50	45.46	46.05	45.59	**57.14**	45.52	47.30	44.16
Sequence	**54.08**	47.62	47.73	46.46	**48.94**	48.89	49.36	46.41	**55.10**	44.90	45.21	42.49	54.24	46.37	46.63	43.35
Emotion Recognition – MELD (7)																
Baseline	43.65	27.84	27.02	27.42	39.64	24.87	24.02	24.43	35.74	20.11	18.43	19.76	41.12	26.08	25.34	25.70
Sequence	**53.32**	38.44	40.08	39.25	**53.72**	39.54	41.32	40.42	**48.32**	32.44	35.71	34.03	**51.26**	24.15	25.56	24.85
Emotion Recognition – IEMOCAP (5)																
Baseline	51.35	54.19	50.63	50.89	39.32	39.68	41.40	40.18	48.11	52.01	48.98	48.60	36.35	40.11	35.02	31.20
Sequence	**58.92**	61.29	58.74	59.32	**53.44**	60.95	59.88	59.76	**59.46**	60.29	58.98	59.49	**49.32**	49.84	50.47	49.74
Sarcasm Detection – MUStARD (2)																
Baseline	58.62	58.65	58.56	58.48	47.06	42.96	48.87	45.72	37.94	35.85	70.56	47.56	41.92	42.01	41.83	41.56
Sequence	**78.16**	80.11	78.01	77.74	**75.38**	69.70	79.31	74.19	**70.67**	67.19	70.49	68.80	**75.86**	78.42	75.69	75.22
Sarcasm Detection – MuSaG (2)																
Baseline	42.43	41.23	41.06	40.91	44.87	42.11	39.95	38.44	41.43	41.03	41.01	42.12	39.57	39.15	39.06	38.83
Sequence	**55.56**	54.65	54.58	54.58	**60.13**	55.78	55.50	55.21	**53.80**	53.38	53.39	53.36	**52.63**	49.72	49.72	49.52
Humor Detection – UR-FUNNY (2)																
Baseline	52.31	52.39	52.31	50.48	**62.88**	66.33	63.19	69.38	49.60	49.79	49.81	48.53	49.30	24.30	49.30	32.55
Sequence	**58.65**	58.74	58.65	58.62	62.47	63.79	62.47	61.74	**66.50**	66.53	66.52	66.50	**60.97**	62.43	61.19	60.06
Humor Detection – M2H2 (2)																
Baseline	58.04	56.09	58.04	56.87	55.45	55.06	55.45	55.25	55.88	55.71	55.88	55.79	58.36	51.41	51.26	51.11
Sequence	**58.35**	55.12	58.36	56.22	**61.71**	55.00	61.71	54.81	**61.38**	56.52	61.38	57.48	**66.49**	59.38	55.18	53.86
Hate Speech Detection – HateMM (2)																
Baseline	82.13	83.82	81.22	81.23	63.73	63.11	63.02	63.00	82.67	84.07	81.21	81.54	76.27	80.64	76.18	74.66
Sequence	**87.33**	86.71	86.80	86.71	**77.33**	78.00	75.44	75.95	**83.20**	83.29	83.35	82.98	**85.33**	85.79	84.89	84.87
Hate Speech Detection – MultiHateClip (English) (2)																
Baseline	**70.60**	66.20	65.85	65.71	60.27	53.44	53.64	52.80	69.13	67.83	61.16	59.83	66.85	65.18	58.60	55.37
Sequence	69.80	60.28	59.79	58.38	**65.24**	57.24	55.22	54.90	**71.01**	56.89	56.97	53.26	**71.41**	67.19	63.82	64.42
Deception Detection – RLTD (2)																
Baseline	52.31	51.87	53.57	47.55	**69.33**	69.29	68.86	68.69	47.69	42.06	48.57	39.07	55.38	58.42	56.43	52.71
Sequence	**66.15**	68.14	65.95	64.06	66.15	66.77	65.48	65.24	**58.46**	62.61	59.05	51.05	**63.08**	66.06	61.90	58.44
Deception Detection – MU3D (2)																
Baseline	53.12	43.26	53.12	46.14	50.00	49.64	50.00	49.23	54.37	50.87	54.37	48.65	50.00	50.93	50.00	47.51
Sequence	**56.87**	54.01	56.87	51.91	**51.88**	52.31	51.88	50.67	**61.54**	65.28	63.10	60.61	**54.06**	53.92	54.06	52.00



(a) UR-FUNNY (Humor)

(b) M2H2 (Humor)

(c) HateMM (Hate Speech)

(d) MultiHateClip (Hate Speech)

(e) RLTD (Deception)

(f) MU3D (Deception)

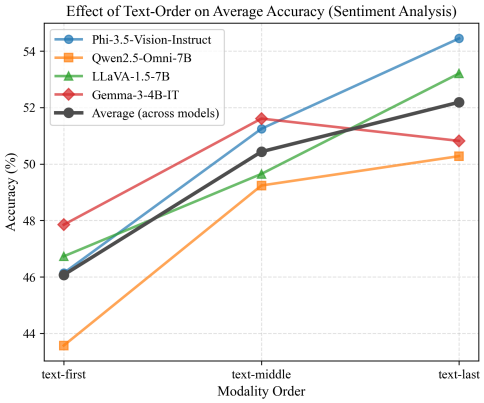
Figure 4: Effect of modality injection order on accuracy across humor, hate speech, and deception detection tasks. Each subplot reports accuracy trends under different modality orders for four backbone models.

Table 10: Relative performance comparison under different modality injection orders. For each order category, accuracy is averaged over all corresponding permutations and datasets. Classification accuracy (%) is reported for six downstream tasks across four multimodal backbone models.

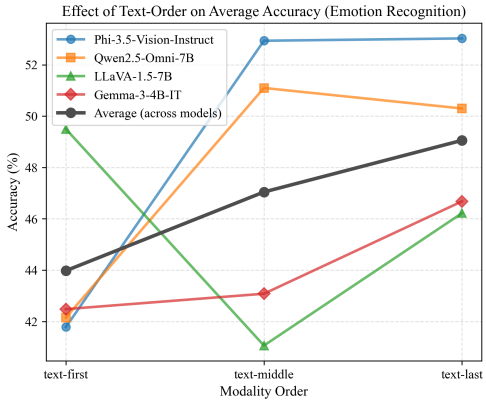
Order	Sentiment Analysis				Emotion Recognition				Sarcasm Detection			
	Phi	Qwen	LLaVA	Gemma	Phi	Qwen	LLaVA	Gemma	Phi	Qwen	LLaVA	Gemma
t-first	46.14	43.57	46.73	47.85	41.78	42.15	49.50	42.48	65.27	62.09	57.21	54.17
t-mid	51.25	49.24	49.65	51.61	52.94	51.10	41.06	43.08	61.81	61.61	53.15	57.35
t-last	54.45	50.28	53.21	50.82	53.03	50.30	46.22	46.68	64.55	63.12	55.89	57.14
v-first	51.99	50.07	51.09	51.34	51.68	49.50	38.28	43.68	62.80	60.21	57.39	59.60
v-mid	51.64	47.49	50.51	50.27	46.93	45.44	52.32	43.32	65.13	61.32	53.76	54.74
v-last	48.22	45.53	48.00	48.67	49.14	48.62	46.18	45.24	63.69	65.30	55.10	54.32
a-first	53.72	49.45	51.77	51.08	54.29	51.90	49.00	46.08	63.55	64.53	51.65	54.88
a-mid	48.95	46.36	49.44	48.40	47.88	47.02	43.40	45.84	64.69	63.89	59.34	56.57
a-last	49.18	47.28	48.38	50.79	45.58	44.63	44.38	40.32	63.38	58.41	55.26	57.20
Order	**Humor Detection**				**Hate Speech Detection**				**Deception Detection**			
	Phi	Qwen	LLaVA	Gemma	Phi	Qwen	LLaVA	Gemma	Phi	Qwen	LLaVA	Gemma
t-first	57.82	57.50	60.26	62.61	77.80	70.05	73.99	75.69	59.27	54.63	57.51	56.79
t-mid	55.92	54.35	56.59	61.98	75.70	68.42	73.13	72.21	55.09	52.95	56.54	55.37
t-last	55.39	56.13	60.51	62.82	74.92	70.95	75.53	76.89	54.54	49.56	56.76	55.10
v-first	55.92	54.83	60.22	62.20	75.12	69.75	74.16	74.40	53.86	46.54	55.39	52.56
v-mid	56.15	59.60	60.07	62.20	75.99	70.25	75.70	75.77	57.26	52.39	55.42	56.56
v-last	57.06	53.56	57.06	63.02	77.30	69.42	72.79	74.63	57.78	58.20	60.00	58.15
a-first	55.39	55.66	56.88	62.60	75.49	69.62	74.50	74.70	55.77	55.96	57.91	57.92
a-mid	57.05	54.03	60.70	63.23	76.72	70.75	73.82	76.82	56.55	51.79	58.85	55.33
a-last	56.68	58.29	59.78	61.58	76.20	69.05	74.33	73.27	56.58	49.38	54.05	54.01

Table 11: Relative performance comparison under different modality injection orders, based on Table 10. The symbols **H**/**M**/**L** indicate high/medium/low performance, determined by relative ranking amongst the same modality perspective for each setting.

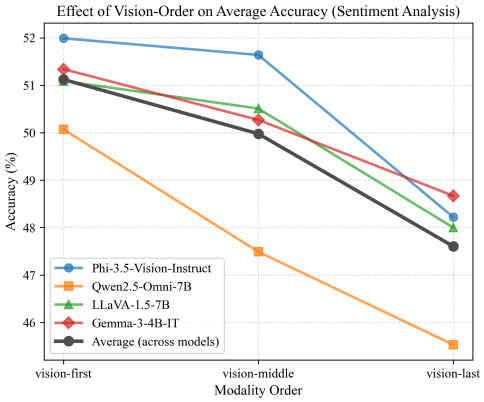
Order	Sentiment Analysis				Emotion Recognition				Sarcasm Detection			
	Phi	Qwen	LLaVA	Gemma	Phi	Qwen	LLaVA	Gemma	Phi	Qwen	LLaVA	Gemma
t-first	L	L	L	L	L	L	H	L	H	M	H	L
t-mid	M	M	M	H	M	H	L	M	L	L	L	H
t-last	H	H	H	M	H	M	M	H	M	H	M	M
v-first	H	H	H	H	H	H	L	M	L	L	H	H
v-mid	M	M	M	M	L	L	H	L	H	M	L	M
v-last	L	L	L	L	M	M	M	H	M	H	M	L
a-first	H	H	H	H	H	H	H	H	M	H	L	L
a-mid	L	L	M	L	M	M	L	M	H	M	H	M
a-last	M	M	L	M	L	L	M	L	L	L	M	H
Order	**Humor Detection**				**Hate Speech Detection**				**Deception Detection**			
	Phi	Qwen	LLaVA	Gemma	Phi	Qwen	LLaVA	Gemma	Phi	Qwen	LLaVA	Gemma
t-first	H	H	M	M	H	M	M	M	H	H	H	H
t-mid	M	L	L	L	M	L	L	L	M	M	L	M
t-last	L	M	H	H	L	H	H	H	L	L	M	L
v-first	L	M	H	M	L	M	M	L	L	L	L	L
v-mid	M	H	M	M	M	H	H	H	M	M	M	M
v-last	H	L	L	H	H	L	L	M	H	H	H	H
a-first	L	M	L	M	L	M	H	M	L	H	M	H
a-mid	H	L	H	H	H	H	L	H	M	M	H	M
a-last	M	H	M	L	M	L	M	L	H	L	L	L



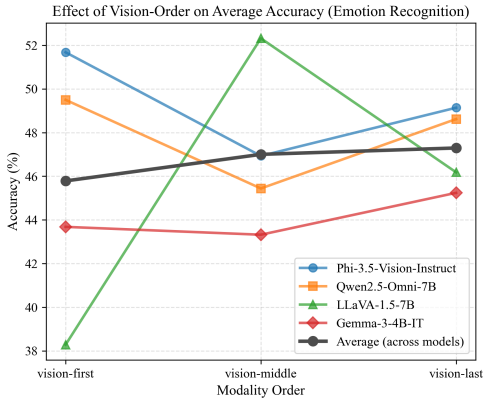
(a) Text–Sentiment



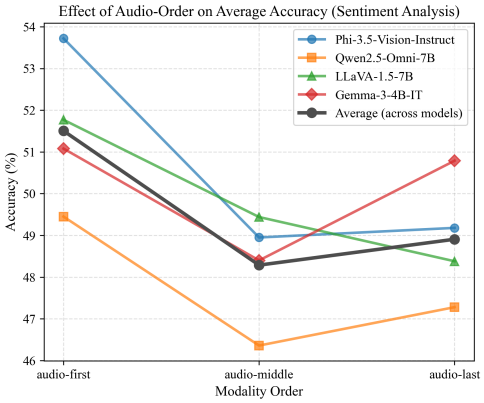
(b) Text–Emotion



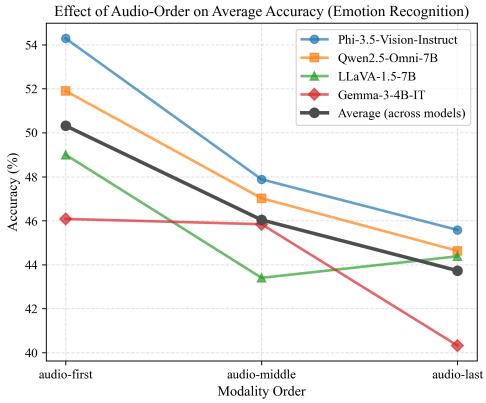
(c) Vision–Sentiment



(d) Vision–Emotion

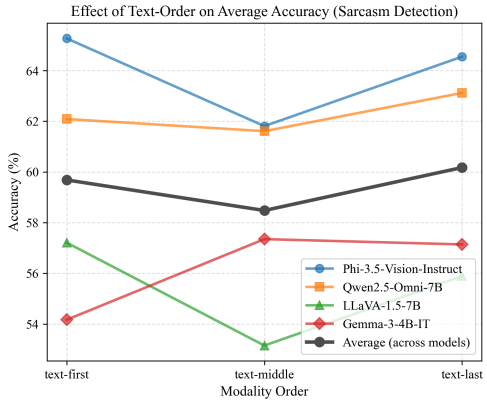


(e) Audio–Sentiment

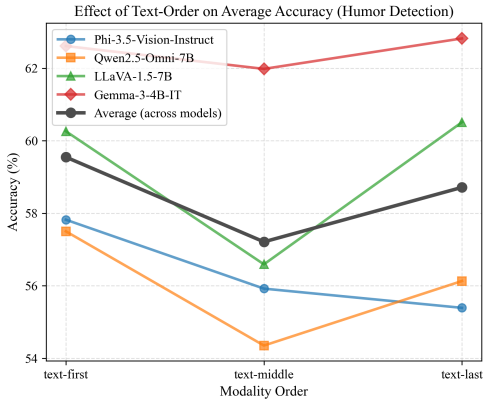


(f) Audio–Emotion

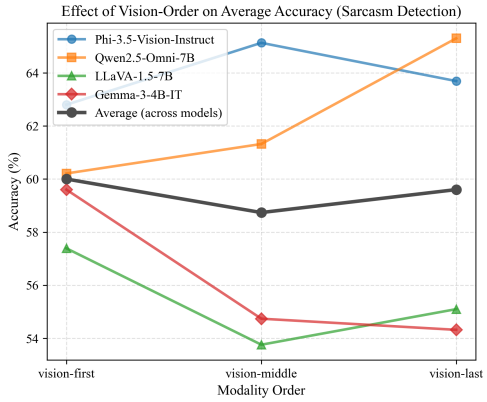
Figure 5: Effect of modality injection order on accuracy for sentiment analysis and emotion recognition across four multimodal backbone models.



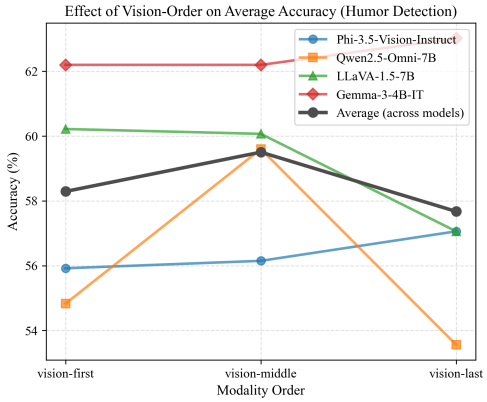
(a) Text–Sarcasm



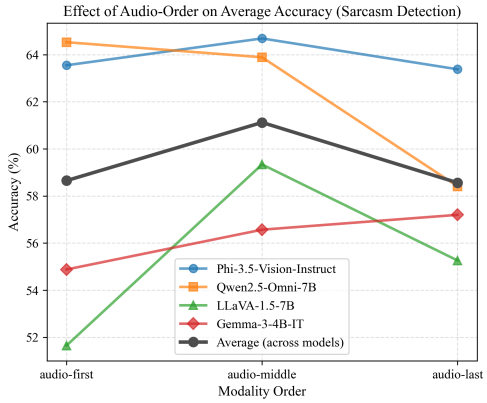
(b) Text–Humor



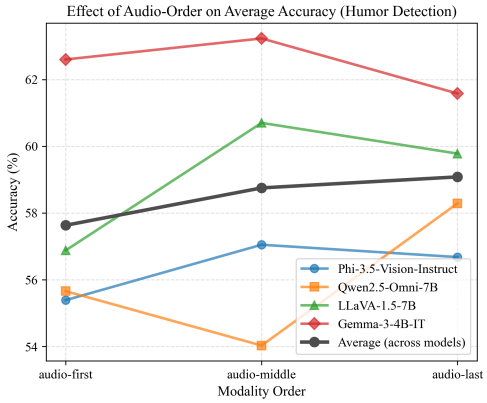
(c) Vision–Sarcasm



(d) Vision–Humor

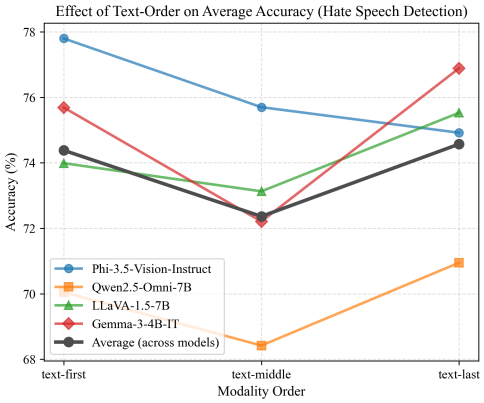


(e) Audio–Sarcasm

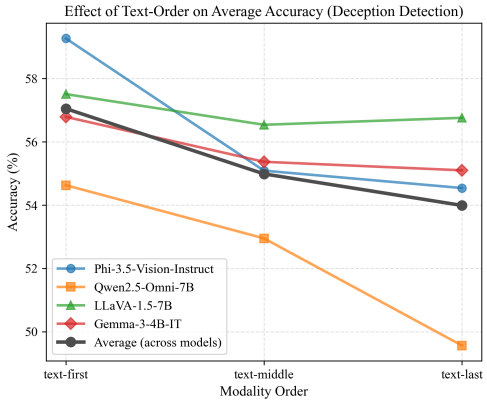


(f) Audio–Humor

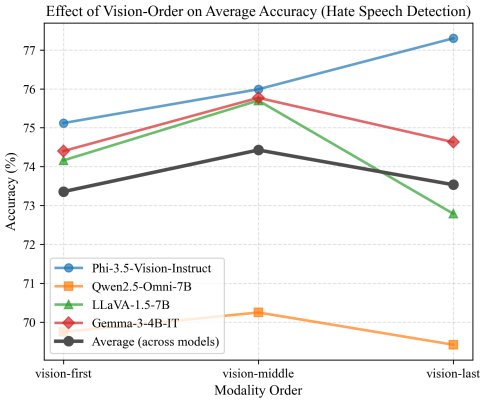
Figure 6: Effect of modality injection order on accuracy for sarcasm detection and humor detection across four multimodal backbone models.



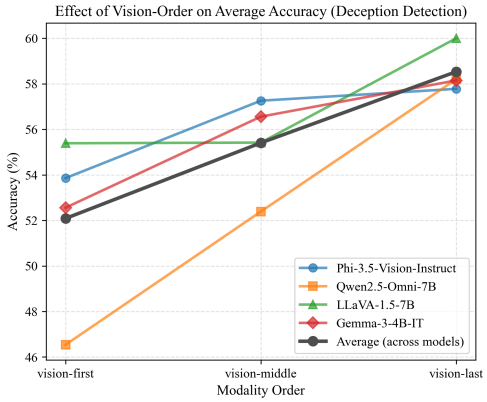
(a) Text–Hate Speech



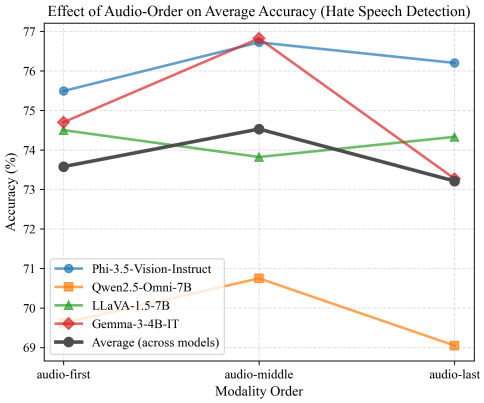
(b) Text–Deception



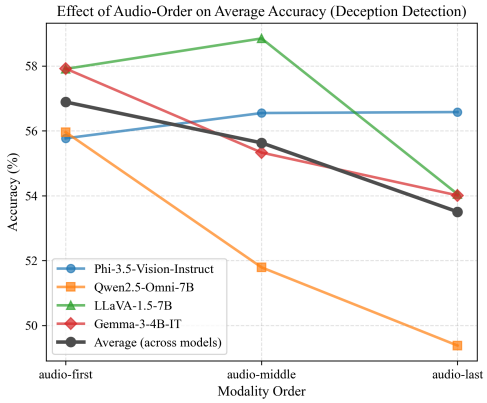
(c) Vision–Hate Speech



(d) Vision–Deception

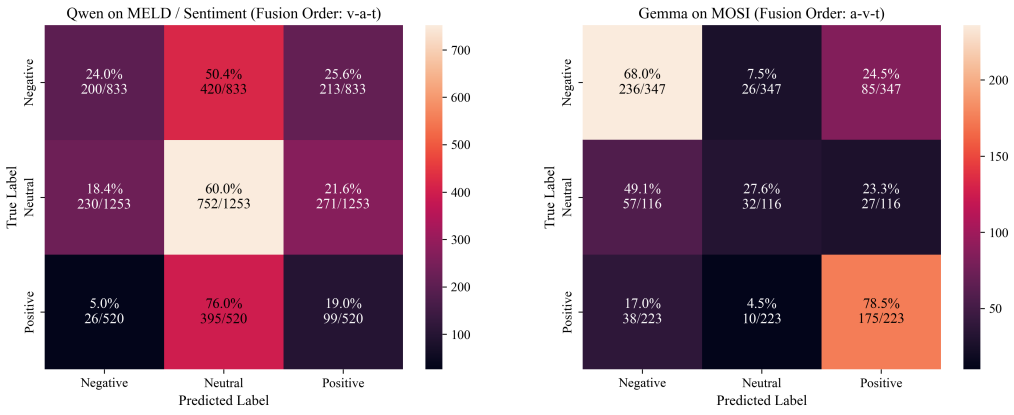


(e) Audio–Hate Speech



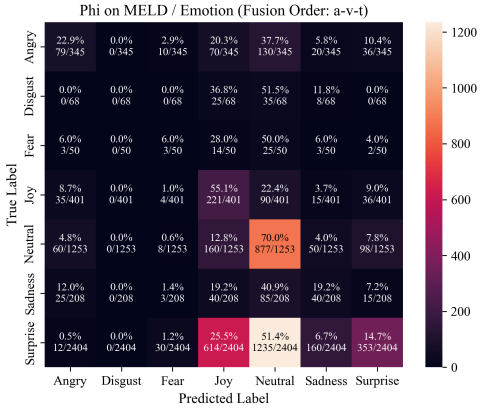
(f) Audio–Deception

Figure 7: Effect of modality injection order on accuracy for hate speech detection and deception detection across four multimodal backbone models.

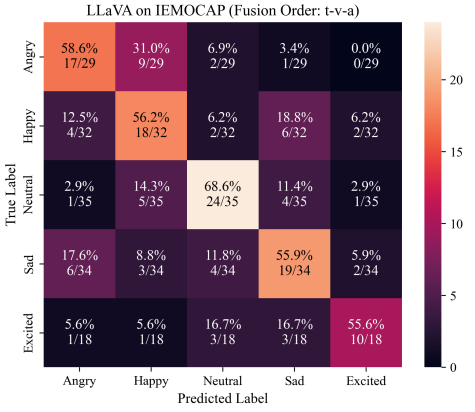


(a) MELD / Sentiment (Qwen, v-a-t)

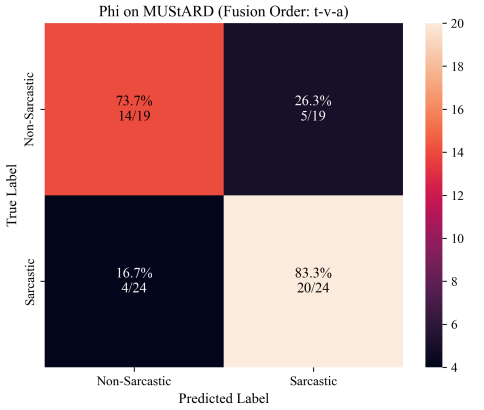
(b) CMU-MOSI / Sentiment (Gemma, a-v-t)



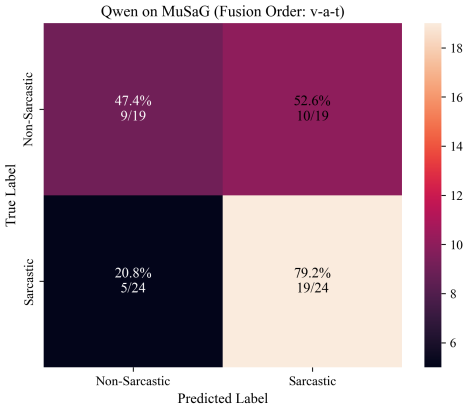
(c) MELD / Emotion (Phi, a-v-t)



(d) IEMOCAP / Emotion (LLaVA, t-v-a)

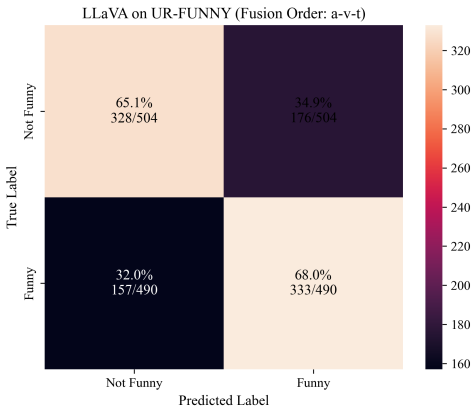


(e) MUStARD / Sarcasm (Phi, t-v-a)

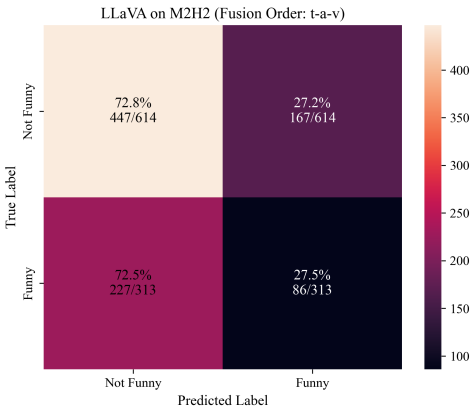


(f) MuSaG / Sarcasm (Qwen, v-a-t)

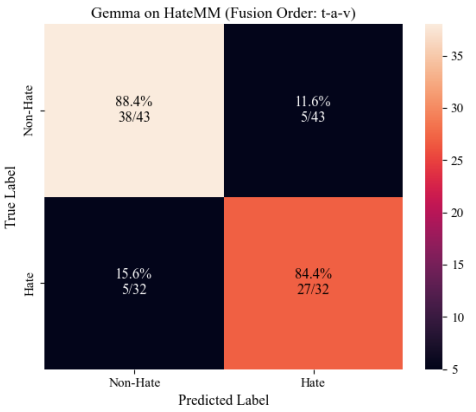
Figure 8: Representative confusion matrices across sentiment, emotion, and humor recognition tasks. Each subplot corresponds to a selected backbone model and modality injection order.



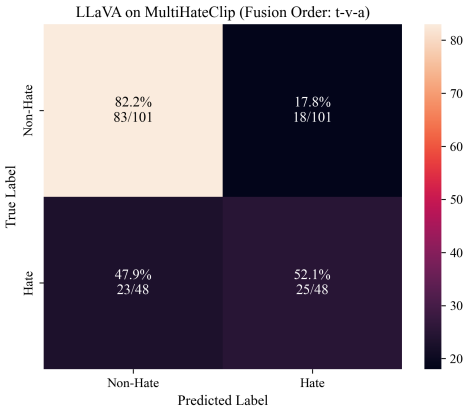
(a) UR-FUNNY / Humor (LLaVA, a-v-t)



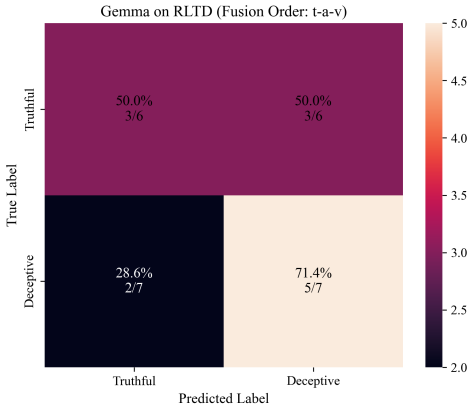
(b) M2H2 / Humor (Phi, a-t-v)



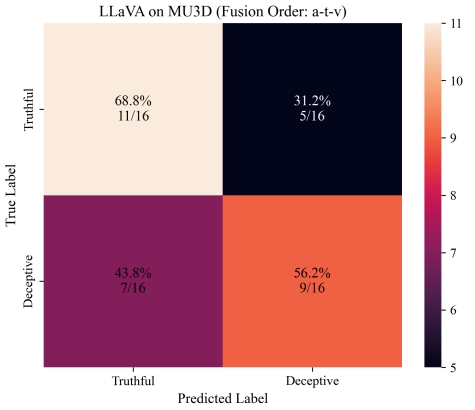
(c) HateMM / Hate Speech (Gemma, t-a-v)



(d) MultiHateClip / Hate Speech (LLaVA, t-v-a)



(e) RLTD / Deception (Gemma, t-a-v)



(f) MU3D / Deception (LLaVA, a-t-v)

Figure 9: Representative confusion matrices across humor, hate speech, and deception detection tasks. Each subplot corresponds to a selected backbone model and modality injection order.