CARES: How Can Missing-Modality Completion Truly Help Multimodal LLMs?

Anonymous authors
Paper under double-blind review

Abstract

Multimodal large language models (MLLMs) have advanced rapidly in recent years. However, their performance can degrade when input modalities are missing. Recent work has used modality completion to improve the robustness of MLLM inference. However, the completed modality is not always beneficial and can even be misleading, since modality completion may introduce inaccurate or irrelevant information. This raises a question: how can missing-modality completion truly help MLLMs? To this end, we propose Completion-Aware Routing and Evidence Scaling (CARES), a lightweight framework with two modules: completion-aware routing dynamically estimates the contribution of completed modalities at the token level, while evidence scaling adaptively regulates their influence during reasoning. Experiments across four benchmarks and multiple MLLM backbones validate the effectiveness, versatility, and generalization of the proposed CARES.

1 Introduction

Multimodal large language models (MLLMs) have advanced vision-language understanding (Liu et al., 2023; Li et al., 2023) and audio-language understanding (Tang et al., 2024; Xu et al., 2025), enabling a wide range of applications such as multimodal question answering, captioning, and sentiment or intent recognition (Pipoli et al., 2025; Zhang et al., 2025). However, in real-world applications, task-relevant modalities may be unavailable at inference time because of sensor failures or transmission errors (Pipoli et al., 2025; Wu et al., 2026). Their absence can degrade MLLM performance (Pipoli et al., 2025). Therefore, robust inference under missing-modality conditions is important for practical deployment.

To handle missing modalities, existing methods can be broadly grouped into three categories: training-based, prompt-based, and completion-based methods. Prompt-based methods guide MLLMs to use the available modalities without updating model parameters (Pipoli et al., 2025). However, simple prompts may yield limited performance gains when task-relevant information is missing. Training-based methods offer stronger adaptation by learning to operate on incomplete inputs (Ma et al., 2021; 2022; Lin & Hu, 2023). Despite achieving better performance, fully fine-tuning billion-parameter MLLMs can be computationally expensive (Hu et al., 2022; Pipoli et al., 2025). Completion-based avoid extensive MLLM fine-tuning. Specifically, they compensate for missing modalities by retrieving relevant content from external sources or generating substitute content conditioned on the available modalities, such as MissRAG (Pipoli et al., 2025).

However, modality completion is not always beneficial and can even interfere with downstream reasoning. As illustrated in Fig. 1(a), the retrieved text may contain erroneous or irrelevant information, which can mislead the model into incorrect predictions. As shown in Fig. 1(b), MissRAG that directly incorporates completed evidence underperforms the available-only baseline in both missing-modality settings. This suggests that the contribution of completed evidence should be adaptively determined by its reliability. To this end, MissRAG uses source-labeling prompts to identify the retrieved modality and lets the MLLM determine its contribution during reasoning. However, source identity alone cannot distinguish useful completion from erroneous or irrelevant content. Consequently, the model may still misjudge the contribution of completed evidence, resulting in limited or even degraded performance (see Fig. 1(b)).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

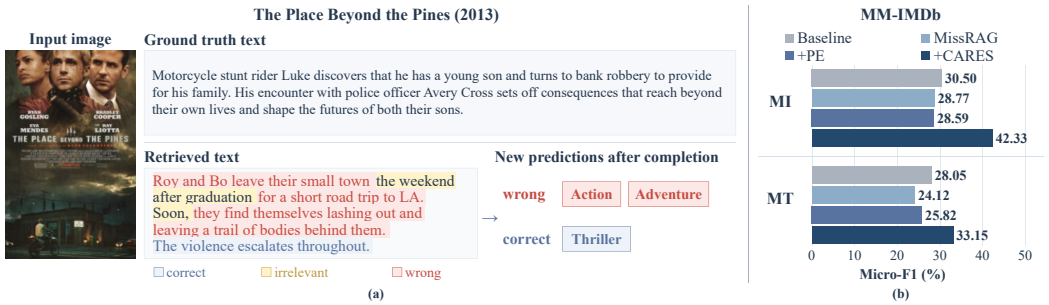


Figure 1: (a) Image-to-text completion for a movie example, the resulting genre prediction shifts. (b) Micro-F1 (%) on MM-IMDb under missing-image (MI) and missing-text (MT) settings, comparing Baseline, MissRAG, MissRAG+PE, and MissRAG+CARES. Both panels use Gemma4 E2B.

To address this limitation, we propose *Completion-Aware Routing and Evidence Scaling* (CARES), a lightweight framework that dynamically regulates the influence of completed evidence during MLLM inference. CARES treats completed modalities as auxiliary evidence whose contribution should be determined adaptively rather than accepted uniformly. Specifically, CARES combines a state-conditioned gate with attention reweighting for completed modalities. At every query token in each layer and head, the gate uses the current attention state to predict a scaling coefficient for each completed modality. CARES then uses this coefficient to adjust how much attention the query assigns to completed modalities while leaving the available inputs unchanged. Only the shared gate is trained, requiring a few thousand parameters while keeping the MLLM backbone frozen.

Our main contributions are summarized as follows:

- We reveal that modality completion is not always beneficial and can even be misleading. This highlights the importance to dynamically regulate the influence of completed evidence according to the model's reasoning needs.
- We propose CARES, a lightweight framework that dynamically regulates completed evidence at the token level to reduce noise interference in MLLM inference.
- Extensive experiments demonstrate the effectiveness of CARES across diverse MLLMs, datasets, and completion methods.

2 Related Work

Learning with missing modalities. Existing approaches improve robustness to incomplete inputs through missing-modality training, representation learning, and parameter-efficient adaptation. Studies of multimodal Transformers (Ma et al., 2022) examine their sensitivity to missing modalities and show that fusion design affects robustness, motivating the learning of dataset-specific fusion strategies. MissModal (Lin & Hu, 2023) aligns complete and incomplete multimodal representations through geometric, distributional, and sentiment-semantic constraints for robust sentiment prediction. MMANet (Wei et al., 2023) combines margin-aware distillation with modality-aware regularization to transfer multimodal knowledge and improve learning for weak modality combinations. DMRNet (Wei et al., 2024) decouples training and inference representations through probabilistic embeddings, preserving modality-specific information across different input combinations. Learnable missing-aware prompts (Lee et al., 2023) provide a parameter-efficient way to adapt pretrained multimodal Transformers to different modality-availability patterns. AOEPT (Lang et al., 2026) distills global modality-specific priors into modal-contextualized prompts and conditions their instantiation on the available modalities to provide instance-specific cues for missing modalities. SimMLM (Li et al., 2025) dynamically weights modality experts within task-specific multimodal models. GIML (Meng & Wei, 2026) jointly addresses missing modalities and corrupted inputs through noise-semantic decoupling and quality-aware adaptive fusion. DyMo (Du et al., 2026) decides whether to incorporate each recovered modality within a task-specific multimodal classification framework. Its selection makes hard inclusion-or-exclusion decisions at the whole-modality level, without explicitly regulating the influence of accepted content at finer computational

positions. These studies address incomplete multimodal learning through representation training, distillation, or adaptive fusion. However, their reported formulations and evaluations only focus on multimodal prediction networks rather than missing-modality inference in MLLMs.

Missing-modality completion. Completion supplements available observations through feature recovery, retrieval, or cross-modal generation. MMIN (Zhao et al., 2021) learns joint multimodal representations to predict missing-modality representations from available inputs for emotion recognition. For multimodal Transformers, RAGPT (Lang et al., 2025) retrieves related instances, reconstructs missing-modality representations with a learnable filter, and generates context-aware dynamic prompts. For MLLMs, MissRAG (Pipoli et al., 2025) retrieves missing-modality content from prototype pools and combines retrieval with modality-aware textual prompts without additional backbone fine-tuning. Generative models offer another way to supply content conditioned on the available modalities. In our experiments, we implement this route using NExT-GPT (Wu et al., 2024), an any-to-any model that supports generation across text, images, video, and audio. CARES complements these approaches by keeping the supplied completion fixed and learning how a frozen MLLM allocates attention between available and completed sources.

Attention regulation in MLLMs. Prior work controls visual information flow through both learned architectural components and inference-time attention interventions. Flamingo (Alayrac et al., 2022) introduces visual information through newly trained gated cross-attention blocks interleaved with frozen language-model layers. PAI (Liu et al., 2025) amplifies attention to image tokens and refines output logits using a text-only contrast to mitigate text inertia and hallucination without additional training. VAR (Kang et al., 2025) analyzes visual attention sinks, selects image-centric heads, and redistributes attention from sink tokens to visual non-sink tokens within those heads. PAI and VAR demonstrate the utility of inference-time attention intervention without parameter updates. While these methods focus on integrating or using visual information, CARES distinguishes evidence from available modalities from content supplied through modality completion. It learns state-conditioned attention allocation between these sources within a frozen MLLM, while preserving relative token attention weights within each source at the current attention operation.

3 METHOD

3.1 OVERVIEW

Most existing MLLMs treat available and completed modalities alike as input evidence. However, completed evidence can contain both useful and misleading information, and different queries may attend to different parts of it. This motivates applying different degrees of attenuation to completed evidence for each queries. To achieve this, we propose *Completion-Aware Routing and Evidence Scaling* (CARES), a lightweight token-level framework integrated into the attention computation of a frozen MLLM. As illustrated in Figure 2, CARES adaptively scales attention weights assigned to completed modalities based on the current attention state.

Notation. We consider missing-modality inference, where an MLLM receives the available inputs together with content supplied by modality completion. Let M denote the number of completed modalities. For a given query, we partition the visible tokens into an available source O and completed sources $C_1, \ldots, C_M$. Here, C_m denotes the set of tokens belonging to the m-th completed modality, while O contains all remaining visible tokens, including available inputs and prompt tokens. At each decoder layer, attention head, and query position, CARES predicts a scaling coefficient r_m for each completed source C_m. For clarity, we omit the layer, head, and query indices unless otherwise specified. We next describe coefficient prediction and attention reweighting.

3.2 STATE-CONDITIONED SHARED GATE

A key question in attenuating completed evidence is how to determine the appropriate attenuation strength. Simply labeling completed modalities or applying a uniform attenuation coefficient across a sample cannot fully address this issue. This is because queries at different layers, attention heads, and positions may attend to different parts of a completed source, which means that the attention allocated to useful and misleading content varies across queries. Therefore, The attenuation applied to completed evidence should depend on the current query. To enable this fine-grained control,

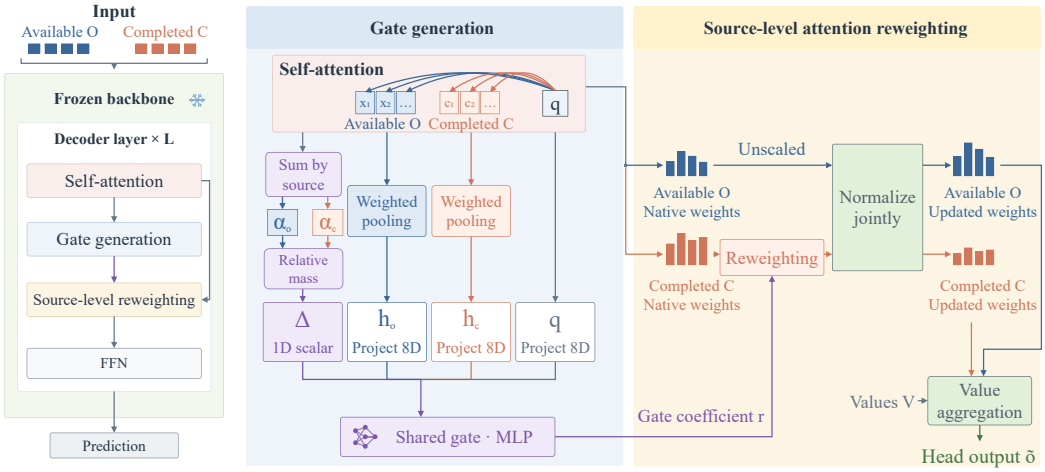


Figure 2: Overview of CARES in a frozen MLLM. At every query token in each decoder layer and attention head, the shared, state-conditioned gate predicts a scaling coefficient for each completed modality. Source-level attention reweighting then adjusts the attention allocated between the available source and completed sources.

CARES first predicts a scaling coefficient r_m for each completed source at each decoder layer, attention head, and query position.

Gate inputs. Consider one sample, decoder layer, attention head, and query position; we omit these indices below for clarity. Let $\mathbf{q}, \mathbf{k}_j, \mathbf{v}_j \in \mathbb{R}^{d_h}$ denote the query, key, and value vectors, and let $\mathcal{V}$ contain the token indices visible to the query. For standard attention (Vaswani et al., 2017), the attention score s_j and normalized attention weight a_j for the j-th visible token are computed as

$$s_j = \frac{\mathbf{q}^\top \mathbf{k}_j}{\sqrt{d_h}}, \qquad a_j = \frac{\exp(s_j)}{\sum_{k\in\mathcal{V}} \exp(s_k)}. \tag{1}$$

CARES retains each backbone's native attention scoring and visibility rules, including any model-specific scaling or bias. For each nonempty visible source $S \in \{O, C_1, \ldots, C_M\}$, we define its attention mass and source-specific representation as

$$\alpha_S = \sum_{j\in S} a_j, \qquad \mathbf{h}_S = \sum_{j\in S} \frac{a_j}{\alpha_S} \mathbf{v}_j. \tag{2}$$

Here, α_S is the source attention mass, and $\mathbf{h}_S$ is its attention-weighted representation. Sources with no visible tokens receive zero attention and are excluded from gating.

Coefficient prediction. For each completed source C_m, the gate receives $\mathbf{q}$, $\mathbf{h}_O$, $\mathbf{h}_{C_m}$, and

$$\Delta_m = \log \frac{\alpha_{C_m}}{\alpha_O}. \tag{3}$$

The query and source representations describe the current computational state, while Δ_m captures the completed source's attention mass relative to the available source before reweighting. We compute this log-ratio from source log-sum-exp values for numerical stability. The gate predicts

$$z_m = G_\theta\left(\mathbf{q}, \mathbf{h}_O, \mathbf{h}_{C_m}, \Delta_m\right), \qquad r_m = \exp\left[-\lambda \sigma(z_m)\right], \tag{4}$$

where σ is the sigmoid function and $\lambda > 0$ controls the attenuation range, ensuring $0 < r_m < 1$. Completed modalities are derived from available inputs rather than directly observed, and may introduce approximation errors. We therefore adopt a conservative correction that attenuates their influence relative to evidence from available modalities, while coefficients close to one retain their native influence.

The gate independently projects each vector input to eight dimensions with a linear layer and SiLU activation. Their concatenation with Δ_m forms a 25-dimensional input to a two-layer MLP with 16

hidden units and a SiLU activation. The gate parameter count scales linearly with the backbone's per-head dimension d_h. For $d_h = 128$, the gate contains 3,529 trainable parameters. Its parameters are shared across samples, layers, heads, query positions, and completed sources, while its outputs $r_{l,h,t,m}$ depend on the attention state.

3.3 Source-Level Attention Reweighting

After predicting scaling coefficient r_m at every query token, CARES uses them to adjust attention allocation across sources. It multiplies each completed source's attention mass by r_m and renormalizes jointly over all sources, preserving the relative attention weights within each source.

Reweighting and aggregation. Using Equation 2, the native attention output can be written as

$$\mathbf{o} = \alpha_O \mathbf{h}_O + \sum_{m=1}^{M} \alpha_{C_m} \mathbf{h}_{C_m}. \tag{5}$$

CARES replaces the source masses with

$$\widetilde{\alpha}_O = \frac{\alpha_O}{\alpha_O + \sum_{n=1}^{M} r_n \alpha_{C_n}}, \qquad \widetilde{\alpha}_{C_m} = \frac{r_m \alpha_{C_m}}{\alpha_O + \sum_{n=1}^{M} r_n \alpha_{C_n}}. \tag{6}$$

Since $\widetilde{\alpha}_{C_m}/\widetilde{\alpha}_O = r_m \alpha_{C_m}/\alpha_O$, the Gate attenuates each completed source relative to the available source. The coefficient r_m is not an attention probability; the normalized mass also depends on competition with the other sources. The routed head output is

$$\widetilde{\mathbf{o}} = \widetilde{\alpha}_O \mathbf{h}_O + \sum_{m=1}^{M} \widetilde{\alpha}_{C_m} \mathbf{h}_{C_m}. \tag{7}$$

The head outputs are then concatenated and passed through the backbone's native output projection.

Preserving within-source routing. Equation 6 is equivalent to adding $\log r_m$ to every attention logit belonging to C_m:

$$\widetilde{s}_j = \begin{cases} s_j, & j \in O, \\ s_j + \log r_m, & j \in C_m. \end{cases} \tag{8}$$

Because all tokens within a source receive the same bias, their relative attention weights remain unchanged at the current attention operation. CARES therefore modifies cross-source allocation without directly altering the query, key, or value vectors. Setting all $r_m = 1$ recovers native attention exactly and defines the corresponding to direct completion without CARES. For multiple completed modalities, their coefficients act within the same normalization, allowing the sources to compete jointly.

4 Experiments

We evaluate whether regulating the use of completed evidence improves missing-modality inference, examine the role of local gating, and analyze routing behavior across computational positions, samples, and text-completion conditions. We further study a larger backbone and transfer across datasets and completion methods.

4.1 Experimental Setup

Datasets and tasks. We evaluate CARES on MOSI (Zadeh et al., 2016), MIntRec2.0 (Zhang et al., 2024), MM-IMDb (Arevalo et al., 2017), and MUSIC-AVQA (Li et al., 2022), covering sentiment classification, intent recognition, multi-label movie genre classification, and audio-visual question answering, respectively. The experimental splits use the official training, validation, and test splits when available. For MOSI and MIntRec2.0, we consider six missing-modality configurations: MA, MV, MT, MAV, MAT, and MVT, where the letters after M identify the missing audio (A), visual (V), and text (T) modalities. MM-IMDb uses missing-image (MI) and missing-text (MT) configurations, while MUSIC-AVQA uses MA and MV. We report classification accuracy on MOSI and MIntRec2.0, answer accuracy on MUSIC-AVQA, and Micro-F1 and Example-F1 on MM-IMDb.

Table 1: Task-level main results (%, higher is better). MissRAG and NExT-GPT denote direct completion without CARES. Bold marks the best available result separately within each provider's three rows, for each backbone and task.

Task	Audio-visual QA		Sentiment analysis							Intent recognition					Multi-label classification	
Dataset	Music-AVQA		MOSI							MIntRec2.0					MM-IMDb	
Metric	Accuracy		Accuracy							Accuracy					Micro-F1 / Example-F1	
Method	MA	MV	MA	MV	MT	MAV	MAT	MVT	MA	MV	MT	MAV	MAT	MVT	MI	MT
Qwen2.5-Omni-7B																
Baseline	44.66	46.84	74.93	73.32	72.45	72.74	31.05	71.28	34.48	30.89	30.99	34.33	8.02	27.40	44.46 / 44.42	49.63 / 50.07
MissRAG	50.33	45.71	78.72	79.01	49.27	79.45	30.76	50.29	27.99	32.42	8.71	28.18	4.33	8.71	40.71 / 40.27	44.05 / 43.88
MissRAG+PE	46.47	45.27	77.41	76.82	50.29	76.82	34.40	50.00	28.87	33.55	8.61	28.92	5.21	9.44	45.92 / 45.45	45.74 / 44.84
MissRAG+CARES	**56.53**	**49.71**	**82.22**	**84.26**	**81.92**	**84.11**	**50.00**	**83.24**	**36.84**	**38.32**	**30.79**	**38.07**	**7.03**	**30.64**	**54.15 / 52.43**	**53.15 / 52.49**
NExT-GPT	47.81	45.92	76.97	78.13	48.25	74.93	40.23	45.48	30.30	31.38	21.54	26.86	7.77	18.99	45.33 / 45.74	45.65 / 45.57
NExT-GPT+PE	46.47	45.27	76.82	76.09	44.17	72.89	40.52	42.27	31.23	31.92	17.81	29.02	7.82	16.63	51.21 / 51.11	45.79 / 44.80
NExT-GPT+CARES	**57.88**	**49.02**	**83.97**	**84.55**	**82.36**	**83.38**	**44.02**	**84.26**	**36.74**	**34.04**	**32.07**	**34.24**	**10.13**	**31.92**	**53.38 / 51.36**	**53.74 / 53.32**
Gemma4 E2B																
Baseline	42.14	31.20	52.48	76.09	29.30	74.78	5.25	64.29	10.72	18.45	10.62	14.31	9.89	13.08	30.50 / 29.88	28.05 / 24.99
MissRAG	42.78	36.87	67.93	74.49	45.63	67.93	21.14	42.86	23.71	34.97	13.43	23.61	5.16	12.49	28.77 / 26.81	24.12 / 23.92
MissRAG+PE	44.23	37.92	69.10	74.93	46.06	74.05	22.45	43.59	27.69	35.91	8.80	30.35	5.12	6.94	28.59 / 26.72	25.82 / 25.50
MissRAG+CARES	**45.02**	**39.81**	**77.84**	**77.41**	**65.89**	**75.51**	**49.85**	**65.16**	**35.07**	**38.17**	**27.25**	**37.04**	**6.98**	**25.58**	**42.33 / 39.36**	**33.15 / 32.07**
NExT-GPT	42.48	23.65	69.10	73.32	48.85	59.48	32.65	36.15	32.81	33.40	23.22	29.66	6.84	21.54	29.43 / 28.80	25.55 / 24.45
NExT-GPT+PE	42.35	25.04	72.01	73.76	44.02	69.24	38.92	45.04	35.51	34.38	16.77	34.04	8.41	17.86	30.88 / 30.56	25.54 / 24.34
NExT-GPT+CARES	**42.79**	**32.58**	**77.26**	**78.72**	**71.14**	**76.97**	**45.19**	**70.41**	**35.71**	**37.92**	**29.41**	**36.06**	**10.67**	**24.40**	**38.77 / 35.53**	**42.35 / 39.89**
MiniCPM-o 4.5																
Baseline	45.19	37.25	79.88	70.26	56.27	79.74	27.41	54.52	35.02	40.04	28.87	38.37	11.07	34.53	51.79 / 50.79	50.74 / 49.66
MissRAG	46.47	38.48	75.36	73.91	41.69	77.70	20.26	47.67	26.76	38.66	11.81	26.36	8.51	11.56	51.37 / 50.31	46.16 / 45.24
MissRAG+PE	48.08	40.23	75.07	70.55	33.53	76.09	15.31	38.34	27.15	38.76	11.61	27.05	9.00	11.41	45.87 / 44.69	44.82 / 44.06
MissRAG+CARES	**56.85**	**49.87**	**85.71**	**85.71**	**84.26**	**84.55**	**52.04**	**83.53**	**37.33**	**45.15**	**33.20**	**37.38**	**9.25**	**38.27**	**53.56 / 52.63**	**50.36 / 49.00**
NExT-GPT	46.26	34.98	71.87	68.51	58.02	74.05	38.19	56.27	34.87	38.42	28.53	35.71	10.82	26.17	50.61 / 49.68	45.91 / 45.01
NExT-GPT+PE	51.12	37.29	74.49	68.66	47.23	75.22	39.94	48.25	34.87	39.74	26.76	35.76	11.17	24.59	50.16 / 49.33	45.90 / 45.05
NExT-GPT+CARES	**62.18**	**47.68**	**86.01**	**85.86**	**83.97**	**85.42**	**47.08**	**85.28**	**41.47**	**45.01**	**41.17**	**40.29**	**12.05**	**42.60**	**52.79 / 50.74**	**49.93 / 48.63**

Backbones and comparison methods. Our main evaluation uses Qwen2.5-Omni-7B (Xu et al., 2025), Gemma 4 E2B (Team et al., 2026), and MiniCPM-o 4.5 (Yao et al., 2024), with retrieval-based completion from MissRAG (Pipoli et al., 2025) and generation-based completion implemented using NExT-GPT (Wu et al., 2024). *Baseline* receives only the available modalities. *MissRAG* and *NExT-GPT* receive the corresponding completed inputs without CARES. For each provider, *+PE* denotes a prompt-only comparison that identifies the completed modalities as approximate inputs, without training or enabling the Gate (Pipoli et al., 2025); *+CARES* applies our source-level regulation. Comparisons within a completion pipeline reuse the same completion content. PE comparisons are retained in the tables, with unreported results marked by dashes.

Implementation and model selection. The multimodal backbone is frozen, and only the shared Gate is optimized. Our reference configuration uses AdamW (Loshchilov & Hutter, 2019) with a learning rate of 10^{-3}, zero weight decay, gradient clipping at 1.0, and an effective batch size of 8. The standard training budget is at most 12 epochs, with validation-based checkpoint selection and early stopping where enabled. Training uses task-specific loss functions: cross-entropy for MOSI, binary cross-entropy for MM-IMDb, answer-token cross-entropy for MUSIC-AVQA, and generalized cross-entropy (Zhang & Sabuncu, 2018) (GCE, $q = 0.7$) for MIntRec2.0. The ablation studies use the same training budget. All experiments are conducted on NVIDIA A800 GPUs.

4.2 Main Results

Table 1 shows that both retrieval-based and generation-based completion can improve available-only inference, but their benefits vary substantially across backbones and missing-modality configurations. In some settings, directly incorporating completed evidence even degrades performance, indicating that obtaining additional input does not ensure its effective use. Adding source-level prompts (+PE) improves some settings but reduces accuracy or F1 in others relative to direct completion, suggesting that textual guidance alone does not reliably improve the use of completed inputs.

By comparison, CARES consistently improves over direct completion across the evaluated backbones, tasks, and completion providers. With MissRAG, CARES improves MOSI accuracy by an average of 18.29 percentage points across the three main backbones and 6 missing-modality configurations. Particularly pronounced gains occur when direct completion degrades available-only

Table 2: Ablations with Qwen and MissRAG. MOSI reports accuracy; MM-IMDb reports Micro-F1 / Example-F1. Param-Matched Tuning (PMT) trains a feature-scaling vector with a similar parameter count comparable to CARES. All scores are percentages (↑). The best result is bold.

	MOSI: Accuracy							MM-IMDb: Micro-F1 / Example-F1		
Method	MA	MV	MT	MAV	MAT	MVT	Avg.	MI	MT	Avg.
Full Gate	**82.22**	**84.26**	**81.92**	**84.11**	**50.00**	**83.24**	**77.63**	**54.15 / 52.43**	**53.15 / 52.49**	**53.65 / 52.46**
Modality	78.72	78.57	53.21	79.45	30.90	52.19	62.17	44.78 / 44.61	48.74 / 48.15	46.76 / 46.38
Sample	78.86	79.01	79.30	79.45	30.90	77.11	70.77	44.82 / 44.63	50.89 / 51.01	47.86 / 47.82
Layer	80.17	79.59	79.30	81.92	46.06	77.70	74.12	49.66 / 48.38	50.50 / 50.34	50.08 / 49.36
Head	81.34	81.78	79.15	82.80	49.27	80.17	75.75	51.25 / 48.81	51.85 / 51.07	51.55 / 49.94
Remove q	80.61	80.61	77.84	82.51	48.40	78.72	74.78	51.00 / 48.85	52.17 / 51.46	51.59 / 50.16
Remove h_O	81.78	82.51	81.49	83.38	47.08	80.32	76.09	52.44 / 50.41	52.50 / 51.65	52.47 / 51.03
Remove h_C	81.34	82.65	80.47	83.38	47.08	78.72	75.61	52.54 / 50.62	51.23 / 50.47	51.89 / 50.55
Remove Δ	81.78	82.65	79.74	82.94	47.38	80.17	75.78	52.07 / 49.47	52.44 / 51.57	52.26 / 50.52
PMT	81.49	81.78	51.75	81.63	46.94	52.19	65.96	35.97 / 33.44	35.99 / 35.13	35.98 / 34.29

performance. In almost all settings, CARES recovers the lost performance and further surpasses the available-only baseline, showing benefits beyond simply reversing completion-induced degradation. Improvements also extend to intent recognition, audio-visual question answering, and multi-label classification, and hold for both retrieval-based and generation-based completion. Since comparisons within each pipeline use identical completed content, these results demonstrate the value of learning how to use completed evidence.

4.3 Ablation Studies

We examine how regulation granularity and Gate inputs contribute to CARES. The former tests where scaling coefficients need to vary, while the latter tests which attention-state signals the Gate needs. Table 2 summarizes both sets of ablations on MOSI and MM-IMDb, covering audio, video, text, and image modalities.

Regulation granularity. We compare four granularity variants: Modality, Sample, Layer, and Head. Each variant is trained separately with the backbone frozen, using the same training budget as Full Gate. Modality learns one input-independent coefficient r_m per completed modality, shared across samples, layers, heads, and query positions. Starting from Sample, the variants use the shared state-conditioned Gate to predict coefficients, with progressively finer output granularity. Sample produces one coefficient $r_{i,m}$ per completed modality per sample, shared across layers, heads, and query positions. Layer additionally allows coefficients to vary across layers, and Head further allows variation across heads. Full Gate additionally allows query-specific coefficients. As illustrate in table 2, sample-dependent regulation improves over learned modality-level scaling in aggregate on both datasets, with a particularly large gain on MOSI. Finer regulation further improves average performance: layer-level regulation yields the largest incremental gain in MM-IMDb Micro-F1, while query-level regulation yields the largest gain in Example-F1. These results support adapting completed-source influence to both the input sample and the local computational state.

Gate inputs. The input ablations examine what information supports this adaptive regulation. The query q represents the current attention query, h_O and h_C summarize the representations read from the available and completed sources, and Δ describes their relative native attention allocation. Removing any one of these inputs reduces aggregate performance on both datasets. Removing q causes the largest drop, while removing either source representation or Δ also degrades performance. These results support conditioning source regulation jointly on the query, the source-specific representations, and their existing attention balance.

Parameter-matched training control. We compare CARES with Param-Matched Tuning (PMT), which trains a feature-scaling vector with a comparable parameter count. Specifically, PMT freezes the Qwen2.5-Omni backbone and learns a 3,584-dimensional element-wise gain applied to the final normalized hidden state before the frozen LM head. Both methods use the same training data, task loss, training budget, and validation-based checkpoint selection protocol. CARES outperforms PMT in all eight evaluated missing-modality settings, with average gains exceeding ten percentage

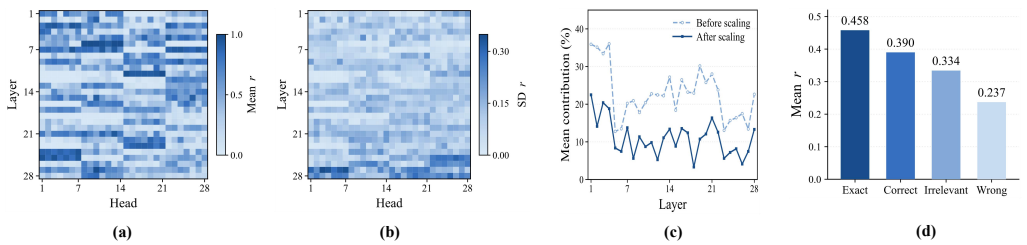


Figure 3: Routing behavior analysis on MOSI with Qwen/MissRAG. **(a, b)** Across-sample mean and standard deviation of r after valid-query averaging; color scales differ. **(c)** Mean completed-source attention mass before and after scaling, averaged across heads within each layer. **(d)** Mean routing coefficient r under four text-completion conditions for the same samples: Exact, Correct, Irrelevant, and Wrong.

points on both datasets. This substantial margin over a similarly supervised control shows that CARES benefits from its state-dependent regulation of completed evidence within attention, beyond the gains from task supervision alone.

4.4 Routing Behavior Analysis

Having established the benefits of state-conditioned regulation, we examine how the learned Gate varies across computational positions and samples, how its coefficients change source attention, and how routing responds to different text-completion conditions. Figure 3 summarizes these analyses. Panels (a)–(c) use the MOSI MT setting with Qwen and MissRAG, covering all test instances. Panel (d) compares mean routing coefficients under four text-completion conditions for the same MOSI samples: Exact, Correct, Irrelevant, and Wrong.

Position-dependent and sample-conditioned regulation. Figure 3(a) shows that the learned Gate assigns different typical regulation strengths to different layer–head positions: after valid-query averaging, the across-sample mean r ranges from approximately 0.010 to 0.988. Figure 3(b) further shows that regulation varies across samples at the same position, with the largest standard deviation reaching 0.349. Thus, the shared Gate learns both position-dependent patterns and sample-conditioned adjustments. These observations complement the granularity ablations: the performance gains from finer regulation are accompanied by learned routing behavior that varies across both positions and inputs.

From coefficients to source attention. The predicted coefficients affect source competition through joint normalization, so their effect also depends on the native attention allocation. Figure 3(c) compares the completed source's attention mass before and after scaling, averaged equally across heads within each layer. Both quantities are measured at the same local attention operation along the Full-Gate trajectory. Averaged equally across all 784 layer–head positions, the completed source's attention mass decreases from 22.79% before scaling to 10.94% after scaling. The layer-wise curves show consistently lower attention mass after scaling, with the magnitude of the reduction varying across layers. This pattern illustrates how the Gate's coefficients translate into changes in actual source allocation.

Routing under different evidence quality. To examine how routing responds to completion quality, we construct four text-completion conditions for the same 600 MOSI target samples. Exact uses the original transcript, Correct uses another transcript with the same sentiment, Irrelevant uses neutral text, and Wrong uses text with the opposite sentiment. The target samples, available inputs, and source labels remain fixed across all four conditions. Figure 3(d) compares the mean routing coefficient r across these conditions. The coefficients follow the ordering Exact (0.458) $>$ Correct (0.390) $>$ Irrelevant (0.334) $>$ Wrong (0.237). This pattern is consistent with weaker attenuation of reliable evidence and stronger attenuation of unrelated or contradictory completions. This shows that CARES adjusts its routing to the content and quality of completed evidence, rather than source identity alone.

4.5 Applicability to a Larger Hybrid Multimodal Backbone

We further evaluate CARES on Nemotron-3-Nano-Omni-30B-A3B, a hybrid multimodal model built on the Nemotron 3 Nano 30B-A3B language backbone (NVIDIA, 2026). Its decoder interleaves 23 Mamba-2 layers, 6 self-attention layers, and 23 MoE layers. We apply source regulation to the six attention layers and train only the shared Gate, keeping the backbone frozen. Table 3 shows consistent improvements over both direct completion and the available-only baseline across all evaluated settings on MOSI and MM-IMDb. Exceeding both references shows that CARES provides benefits beyond recovering completion-induced losses. Notably, regulating only six attention layers improves predictions without modifying the Mamba-2 or MoE components, suggesting that source regulation remains effective within a hybrid decoder. These results also support the portability of CARES to a larger backbone with a hybrid architecture.

Table 3: CARES on Nemotron-3-Nano-Omni-30B-A3B. Scores (%); Δ: gain over MissRAG (pp). For MM-IMDb, Micro and Example denote Micro-F1 and Example-F1.

Task	Base.	MissRAG	CARES	Δ
MOSI: Accuracy				
MA	74.93	70.26	**78.86**	+8.60
MV	66.62	71.57	**79.15**	+7.58
MT	58.02	53.50	**67.93**	+14.43
MAV	72.01	68.95	**82.51**	+13.56
MAT	14.87	19.68	**43.88**	+24.20
MVT	41.69	52.04	**73.03**	+20.99
MM-IMDb: F1				
MI Micro	25.47	8.67	**44.15**	+35.48
MI Example	18.84	6.39	**40.79**	+34.40
MT Micro	13.44	10.23	**23.33**	+13.10
MT Example	10.59	8.11	**18.60**	+10.49

4.6 Cross-Dataset and Cross-Completion Transfer

Transfer settings. We evaluate frozen-Gate transfer to Qwen/MOSI with NExT-GPT completions. Cross-dataset transfer uses Gates trained on MIntRec2.0 with NExT-GPT, while cross-completion transfer uses Gates trained on MOSI with MissRAG. Missing-modality configurations are matched between source and target, with no target-side training. Only the learned gate is transferred.

Cross-dataset transfer. Table 4 shows that Gates learned for intent recognition improve sentiment classification, exceeding both NExT-GPT and the available-only baseline in average accuracy. These cross-dataset gains without target-side training support the view that CARES learns a transferable evidence-regulation policy, rather than merely adapting to the supervised source task.

Cross-completion transfer. Gates trained on retrieval-based completions improve generation-based inference in all configurations. These gains without target-side training suggest that CARES learns transferable, state-dependent scaling of completed evidence that generalizes beyond the training provider.

Table 4: Frozen-Gate transfer to Qwen/MOSI. Dataset transfer fixes NExT-GPT; provider transfer fixes MOSI. Accuracy (%); parentheses show average gains over NExT-GPT (pp).

Task	Baseline	NExT-GPT	MIntRec2.0 $\rightarrow$ MOSI	MissRAG $\rightarrow$ NExT-GPT
MA	74.93	76.97	78.13	80.32
MV	73.32	78.13	77.99	81.92
MT	72.45	48.25	69.83	75.66
MAV	72.74	74.93	76.38	81.20
MAT	31.05	40.23	40.38	40.38
MVT	71.28	45.48	79.15	79.30
Avg.	65.96	60.67	70.31 (+9.64)	73.13 (+12.46)

5 Conclusion

We presented CARES, a lightweight framework that learns how frozen MLLMs use completed evidence through state-conditioned source attention regulation. Experiments across backbones, tasks, and completion providers show that CARES improves predictions without changing the completed evidence. Further results support Gate transfer across datasets and completion providers. These findings support our central insight: MLLMs can derive greater benefit from modality completion by adaptively regulating the influence of completed evidence at the token level.

AI USE STATEMENT

We used large language models to assist with drafting and refining parts of the manuscript, improving grammatical accuracy and readability, and checking selected portions of the code and mathematical formulations. The authors led the conceptual development, methodological design, and analysis, and retained final judgment over all scientific claims. The authors are responsible for verifying the accuracy of AI-assisted content and take full responsibility for the final manuscript and implementation.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. Gated multimodal units for information fusion, 2017. URL https://arxiv.org/abs/1702.01992.

Siyi Du, Xinzhe Luo, Declan P. O'Regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In *International Conference on Learning Representations (ICLR) 2026*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR 2022*, April 2022. URL https://www.microsoft.com/en-us/research/publication/lora-low-rank-adaptation-of-large-language-models/.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/da8a39bc39ae1c89dd6ebb1e3bcbb3f3-Paper-Conference.pdf.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning, 2025. URL https://arxiv.org/abs/2501.01120.

Jian Lang, Rongpei Hong, Ting Zhong, and Fan Zhou. Aoept: Breaking the implicit modality-reduction bottleneck in modality-missing prompt tuning, 2026. URL https://arxiv.org/abs/2605.24816.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14943–14952, June 2023.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19108–19118, June 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Sijie Li, Chen Chen, and Jungong Han. Simmlm: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24068–24077, October 2025.

Ronghao Lin and Haifeng Hu. MissModal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11: 1686–1702, 2023. doi: 10.1162/tacl_a_00628. URL https://aclanthology.org/2023.tacl-1.94/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 125–140, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73010-8.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality, 2021. URL https://arxiv.org/abs/2103.05677.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18177–18186, June 2022.

Xiangyu Meng and Shicai Wei. General incomplete multimodal learning via dynamic quality perception, 2026. URL https://arxiv.org/abs/2607.06943.

NVIDIA. Nemotron 3 nano omni: Efficient and open multimodal intelligence, 2026. URL https://arxiv.org/abs/2604.24954.

Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. Missrag: Addressing the missing modality challenge in multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3215–3224, October 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/476ab8f369e489c04187ba84f68cfa68-Paper-Conference.pdf.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen,

Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20039–20049, June 2023.

Shicai Wei, Yang Luo, Yuji Wang, and Chunbo Luo. Robust multimodal learning via representation decoupling, 2024. URL https://arxiv.org/abs/2407.04458.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey, 2026. URL https://arxiv.org/abs/2409.07825.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53366–53397. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wu24e.html.

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.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

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.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, jinyue Zhao, Wenrui Li, and Yanting Chen. Mintrec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 46367–46393, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/ca97c1a8c52889c49e16497912244c3b-Paper-Conference.pdf.

Hanlei Zhang, zhuohang li, Hua Xu, Yeshuang Zhu, Peiwu Wang, Haige Zhu, Jie Zhou, and Jinchao Zhang. Can large language models help multimodal language analysis? mmla: A comprehensive benchmark. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-3567. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/99e6db87fb42f3c9d2d870984e1319db-Paper-Datasets_and_Benchmarks_Track.pdf.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/f2925f97bc13ad2852a7a551802feea0-Paper.pdf.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2608–2618, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.203. URL https://aclanthology.org/2021.acl-long.203/.

A Mathematical Properties of CARES

This section derives the source-level formulation in Section 3 and its equivalence to an additive attention-logit bias. All statements refer to a single attention operation, with its query, visible keys, values, and predicted coefficients held fixed. Layer, head, and query indices are omitted for readability.

A.1 Source-Level Decomposition

Let $\mathcal{V}$ be the set of keys visible to the current query under the backbone's native attention mask. The sets $O, C_1, \ldots, C_M$ partition $\mathcal{V}$. For native logits s_j, define

$$Z_S = \sum_{j \in S} \exp(s_j), \qquad Z = \sum_{j \in \mathcal{V}} \exp(s_j), \qquad a_j = \frac{\exp(s_j)}{Z}. \tag{9}$$

For a nonempty source S, its mass and normalized representation are

$$\alpha_S = \frac{Z_S}{Z}, \qquad \mathbf{h}_S = \frac{1}{Z_S} \sum_{j \in S} \exp(s_j) \mathbf{v}_j. \tag{10}$$

Consequently,

$$\sum_{j \in \mathcal{V}} a_j \mathbf{v}_j = \sum_{S \in \{O, C_1, \ldots, C_M\}} \alpha_S \mathbf{h}_S. \tag{11}$$

This is an exact decomposition of native attention. It introduces no approximation to token weights within a source.

A.2 Equivalence to Source-Wise Logit Bias

Write $b_O = 0$ and $b_{C_m} = \log r_m$. Adding b_S to every visible logit in source S gives the partition function

$$\widetilde{Z} = Z_O + \sum_{m=1}^{M} r_m Z_{C_m}. \tag{12}$$

With $D = \widetilde{Z}/Z = \alpha_O + \sum_m r_m \alpha_{C_m}$, the resulting token weights satisfy

$$\widetilde{a}_j = \begin{cases} a_j / D, & j \in O, \\ r_m a_j / D, & j \in C_m. \end{cases} \tag{13}$$

Summing Equation 13 within each source recovers Equation 6. Multiplying by the values and summing over all visible keys recovers Equation 7. Thus, source-level aggregation and source-wise logit bias compute the same routed head output.

For any two visible tokens j, k in the same source,

$$\frac{\widetilde{a}_j}{\widetilde{a}_k} = \frac{a_j}{a_k} = \exp(s_j - s_k). \tag{14}$$

Equivalently, $\widetilde{a}_j / \widetilde{\alpha}_S = a_j / \alpha_S$ for $j \in S$. CARES therefore preserves the source-specific representation $\mathbf{h}_S$ at the current attention operation while changing its weight in the final aggregation.

A.3 Bounded Regulation and Joint Source Competition

The parameterization in Equation 4 gives

$$-\lambda < \log r_m < 0, \qquad e^{-\lambda} < r_m < 1. \tag{15}$$

It implements a bounded correction to the backbone's native source allocation. A coefficient near one retains a completed source's native influence relative to evidence from available modalities, while smaller coefficients provide stronger attenuation. The Gate learns this degree of retention from the current query and source representations, rather than assigning one discount to every completion.

For each completed source with positive native mass,

$$\frac{\widetilde{\alpha}_{C_m}}{\widetilde{\alpha}_O} = r_m \frac{\alpha_{C_m}}{\alpha_O}. \tag{16}$$

For two completed sources, joint normalization yields

$$\frac{\widetilde{\alpha}_{C_m}}{\widetilde{\alpha}_{C_n}} = \frac{r_m}{r_n} \frac{\alpha_{C_m}}{\alpha_{C_n}}. \tag{17}$$

Thus, completed sources can receive different relative allocations even though every coefficient lies below one. Their normalized masses depend on all sources participating in the same attention operation.

The total allocation to completed evidence also has a simple form. Let $A_C = \sum_m \alpha_{C_m}$ and $R_C = \sum_m r_m \alpha_{C_m}$. Since $\alpha_O = 1 - A_C$,

$$\widetilde{A}_C = \frac{R_C}{1 - A_C + R_C}. \tag{18}$$

When $0 < A_C < 1$, $R_C < A_C$ implies $\widetilde{A}_C < A_C$. This aggregate property is consistent with source-specific competition: the relative balance among completed sources is still adaptive.

Setting all coefficients externally to one gives $D = 1$ and $\widetilde{a}_j = a_j$, recovering native attention exactly. This corresponds to direct completion without CARES. If no completed source is visible, the attention operation likewise reduces to native attention.

B Architecture and Implementation Details

B.1 Shared Gate Architecture and Initialization

The Gate uses three separate projections to encode the query and the two source representations. Writing ϕ for SiLU,

$$\begin{aligned} \mathbf{u}_m &= [\phi(W_q \mathbf{q} + \mathbf{b}_q); \phi(W_O \mathbf{h}_O + \mathbf{b}_O); \phi(W_C \mathbf{h}_{C_m} + \mathbf{b}_C); \Delta_m], \\ \mathbf{u}_m &\in \mathbb{R}^{25}, \end{aligned} \tag{19}$$

where each vector projection has eight output dimensions. The remaining mapping is

$$z_m = \mathbf{w}_2^\top \phi(W_1 \mathbf{u}_m + \mathbf{b}_1) + b_2, \qquad W_1 \in \mathbb{R}^{16 \times 25}. \tag{20}$$

All linear layers include biases. For a common input head dimension d_h, the trainable parameter count is

$$N_\theta = 3(8d_h + 8) + (25 \times 16 + 16) + (16 + 1) = 24d_h + 457. \tag{21}$$

For Qwen2.5-Omni-7B, $d_h = 128$ gives $N_\theta = 3{,}529$. The count does not grow with the number of decoder layers, attention heads, query positions, or completed sources, because these positions reuse the same Gate parameters.

We fix $\lambda = \log 10000$, giving a coefficient range of $(0.0001, 1)$, and initialize the Gate to $r_0 = 0.5$. Specifically, the final linear weight is initialized to zero and its bias to

$$b_2 = \log \frac{\eta}{1 - \eta}, \qquad \eta = -\frac{\log r_0}{\lambda}. \tag{22}$$

The initial coefficient is consequently independent of the input state; training then learns state-dependent adjustments through the shared network. The backbone, including its modality encoders and LM head, remains frozen.

B.2 Source Assignment and Visibility

Source membership is defined on the processed decoder input. Tokens representing a supplied missing modality belong to its completed source C_m. For completed text, token membership follows the character span of the supplied content after tokenization. For image, video, and audio inputs, membership follows the corresponding modality-content positions produced by the backbone processor. Multiple frames from one completed video belong to the same completed source.

The available source is the complement of the completed sources among visible tokens, as defined in Section 3.1. It includes available modality content and prompt tokens. Task instructions, separators, and modality delimiters outside the marked content positions remain in O. Answer-prefix tokens and subsequently generated answer tokens also belong to O.

At query position t, each source is intersected with the visible key set for that query. A completed source with no visible keys contributes zero mass and is excluded from coefficient prediction. The same source-assignment rule is used during training and evaluation.

Table 5: Source assignment used by CARES. Visibility is additionally restricted by the backbone's native attention mask.

Token content	Source
available text, image, video, or audio content	O
Content supplied for the m-th missing modality	C_m
Task instructions and prompt formatting outside completed spans	O
Answer prefix and generated answer tokens	O
Padding or tokens excluded by the native visibility mask	Not visible

B.3 Stable Source Computation

For each visible source, compute its log-partition value

$$L_S = \mathrm{logsumexp}_{j \in S}(s_j). \tag{23}$$

Then $\Delta_m = L_{C_m} - L_O$, avoiding division of small attention masses. The routed source weights can be computed directly as

$$(\widetilde{\alpha}_O, \widetilde{\alpha}_{C_1}, \ldots, \widetilde{\alpha}_{C_M}) = \mathrm{softmax}(L_O, L_{C_1} + \log r_1, \ldots, L_{C_M} + \log r_M). \tag{24}$$

This form uses the native attention scores, including the backbone's scaling, positional transformations, and visibility restrictions. Source representations are computed using source-normalized attention and combined using these routed masses.

The source-partition implementation can recover the available-side contribution from native attention and the completed-source contributions:

$$\alpha_O = 1 - \sum_m \alpha_{C_m}, \qquad \alpha_O \mathbf{h}_O = \mathbf{o} - \sum_m \alpha_{C_m} \mathbf{h}_{C_m}. \tag{25}$$

The Qwen implementation computes source masses and Gate features in FP32 and uses a 10^{-5} numerical floor when recovering a very small available-side mass. Padded positions do not contribute to source statistics or feature aggregation.

For clarity, a routed attention operation consists of the following steps:

1. Compute the native query, key, and value representations and apply the native positional transformations and visibility mask.
2. Obtain L_S and $\mathbf{h}_S$ for each nonempty visible source.
3. For each completed source, form $\Delta_m = L_{C_m} - L_O$ and predict r_m with the shared Gate.
4. Normalize the source log-partition values with the predicted biases and compute $\widetilde{\mathbf{o}} = \sum_S \widetilde{\alpha}_S \mathbf{h}_S$.
5. Apply the native head concatenation and output projection.

B.4 Task Scoring and Autoregressive Inference

Classification uses the scoring positions specified in Section C.6. The input sequence is processed with causal attention, and the loss is computed from the relevant frozen LM-head outputs. Full Gate regulates valid attention operations throughout the decoder, rather than only the final classification position. Gradients propagate through the frozen decoder to the Gate; only Gate parameters are passed to the optimizer.

MUSIC-AVQA uses answer-token supervision during training. Qwen, MiniCPM, and Gemma use autoregressive answer generation during evaluation, as described in Section C.6. During cached decoding, completed-source positions retain their original membership, and newly generated positions extend O. The current query is evaluated against the visible cached keys using the same source-level regulation. The source masks therefore follow the growing key sequence without reclassifying the original completed evidence.

Table 6: Dataset splits and evaluation tasks. MUSIC-AVQA counts question–answer instances; MM-IMDb is a 23-label classification task. MIntRec2.0 uses the 30 in-scope intent classes.

Dataset	Train	Validation	Test	Reported metric
MOSI	1,284	229	686	Accuracy
MIntRec2.0	6,165	1,106	2,033	Accuracy
MM-IMDb	15,552	2,608	7,799	Micro-F1 / Example-F1
MUSIC-AVQA	31,927	4,568	9,129	Answer accuracy

Table 7: Input availability for the six audio–video–text configurations.

Configuration	available modalities	Completed modalities
MA	Video, text	Audio
MV	Audio, text	Video
MT	Audio, video	Text
MAV	Text	Audio, video
MAT	Video	Audio, text
MVT	Audio	Video, text

B.5 Backbone-Specific Integration

Qwen2.5-Omni-7B uses 28 decoder attention layers with 28 query heads per layer and a head dimension of 128. CARES preserves its grouped-query attention and multimodal rotary positional embeddings. The routing analysis consequently contains $28 \times 28 = 784$ layer–head positions.

For Gemma 4 E2B, the Gate shares a 256-dimensional feature interface across attention layers with different head dimensions. A 256-dimensional vector is used directly. For a 512-dimensional vector $\mathbf{x} = [\mathbf{x}^{(1)}; \mathbf{x}^{(2)}]$, the Gate input is $(\mathbf{x}^{(1)} + \mathbf{x}^{(2)})/\sqrt{2}$. This parameter-free mapping is applied only to Gate features; native attention uses the original vectors and layer-specific visibility rules. The resulting Gate has $24 \times 256 + 457 = 6{,}601$ trainable parameters.

MiniCPM-o 4.5 retains its native multimodal preprocessing and decoder attention transformations. CARES operates on the resulting decoder source positions through the same coefficient-prediction and source-reweighting interface.

For Nemotron-3-Nano-Omni-30B-A3B, CARES is attached only to the six self-attention layers of the hybrid decoder. The 23 Mamba-2 layers and 23 MoE layers retain their native computations. One Gate is shared across the regulated attention layers, and the entire backbone remains frozen.

C Experimental Protocols

C.1 Datasets and Official Splits

We use the official training split for optimization, the official validation split for model selection, and the official test split for evaluation. Table 6 lists the split sizes for the task definitions used in this paper. Missing-modality configurations change the available inputs while retaining the corresponding task annotations and split membership. The datasets are MOSI (Zadeh et al., 2016), MIntRec2.0 (Zhang et al., 2024), MM-IMDb (Arevalo et al., 2017), and MUSIC-AVQA (Li et al., 2022).

C.2 Missing-Modality Configurations

The letters following M specify missing modalities. For MOSI and MIntRec2.0, A, V, and T denote audio, video, and text. Table 7 makes the available and completed inputs explicit. MM-IMDb uses MI and MT, with image and text respectively missing. MUSIC-AVQA uses MA and MV; the task question is provided in both configurations.

Table 8: Reference NExT-GPT generation settings. Video dimensions refer to generated media before downstream backbone preprocessing.

Setting	Value
Checkpoint	`nextgpt_7b_tiva_v0`
Maximum text generation length	160 tokens
Temperature / top-p	0.2 / 1.0
Diffusion steps	25
Guidance scale	7.5
Audio duration	9 seconds
Video resolution / frames	320×576 / 16

available-only baseline receives the available modalities and task instructions. Completion-based methods additionally receive supplied content for each missing modality. Modality content is processed using the corresponding backbone's processor and chat template. In the Qwen implementation, video preprocessing uses up to eight uniformly sampled frames, and audio is decoded separately from the video stream.

C.3 Completion Pipelines and Reuse

We use retrieval-based completion from MissRAG (Pipoli et al., 2025) and generation-based completion from NExT-GPT (Wu et al., 2024). A completion pipeline specifies the dataset, missing-modality configuration, completion provider, and the resulting sample-to-completion mapping. Completed evidence is prepared before downstream Gate training and reused for comparisons within that pipeline. It is not updated when the Gate parameters change.

The direct-completion reference, +PE, and +CARES receive the same completed evidence within a pipeline. The downstream backbone processes that content using its own modality encoders and tokenizer. Thus, source regulation changes how supplied evidence is used, while holding the completion content fixed for the corresponding comparison.

C.4 Retrieval-Based Completion

We follow the ImageBind-based retrieval procedure of MissRAG (Pipoli et al., 2025) to complete missing modalities from the available observations. The retrieval pool consists exclusively of the corresponding official training split, excluding the query itself for training examples; validation and test samples are not retrieval candidates. Each missing modality receives one retrieved completion. The retrieval encoder and completed evidence remain fixed during Gate training, and comparisons within each pipeline reuse the same completions.

C.5 Generation-Based Completion

We use the NExT-GPT `nextgpt_7b_tiva_v0` checkpoint with its multimodal generation components. The generator receives the available modality content and an instruction specifying the desired output modality. The generated text or media is saved before downstream evaluation and remains fixed across the corresponding direct-completion, +PE, and +CARES comparisons.

Table 8 gives the reference generation configuration. Generation saves one completed output per requested modality. Successful outputs are retained when generation resumes; missing or failed outputs are retried before the completion mapping is used downstream. The downstream Gate does not select among generated candidates.

Multiple missing modalities. Across datasets and missing-modality configurations, each missing modality is generated in a single stage conditioned on the original available modalities. For MOSI MAT, both audio and text are generated from the available video; for MVT, both video and text are generated from the available audio. For MAV, audio and video are generated from the available transcript. The same conditioning rule applies to MIntRec2.0 and the other generation-based completion settings.

Generation instructions. For MOSI, the available audio and video are supplied using the native media input format, and available text is introduced by `Transcript: [TEXT]`. The output-specific instructions are:

```
Given the observed [AVAILABLE MODALITIES] of a CMU-MOSI
sentiment sample, write one concise transcript-like
emotional text description for the missing text modality.
Answer with text only.

Given the observed [AVAILABLE MODALITIES] of a CMU-MOSI
sentiment sample, generate one audio clip that matches the
speaker's sentiment and speaking style.  Answer with audio
only.

Given the observed [AVAILABLE MODALITIES] of a CMU-MOSI
sentiment sample, generate one short video clip that matches
the speaker and sentiment.  Answer with video only.
```

MIntRec2.0 uses the corresponding dialogue-utterance instructions:

```
Given only the observed [AVAILABLE MODALITIES] of a dialogue
utterance, reconstruct one concise plausible transcript for
the missing text modality.  Do not infer or output an intent
label.  Answer with text only.

Given only the observed [AVAILABLE MODALITIES] of a dialogue
utterance, generate one plausible speech audio clip for
the missing audio modality.  Do not use any intent label.
Answer with audio only.

Given only the observed [AVAILABLE MODALITIES] of a dialogue
utterance, generate one plausible short visual clip for
the missing video modality.  Do not use any intent label.
Answer with video only.
```

MM-IMDb uses the following instructions, together with the available poster for MT or plot for MI:

```
Given this observed movie poster, write one concise
plot-like synopsis that matches the likely movie genres.
Answer with text only.

Observed movie plot:  [PLOT] Generate one representative
movie poster image that matches this plot and its likely
genres.  Answer with image only.
```

For MUSIC-AVQA, completion uses the available audio or video without the task question or reference answer. Questions associated with the same video reuse its generated completion:

```
Given only the observed video from a Music-AVQA sample,
generate one plausible audio clip for the missing audio
modality.  Do not use any target answer, caption, question,
label, or retrieved prototype.  Answer with audio only.

Given only the observed audio from a Music-AVQA sample,
generate one plausible short video clip for the missing
video modality.  Do not use any target answer, caption,
question, label, or retrieved prototype.  Answer with video
only.
```

C.6 Task Formulation and Output Scoring

All task losses are computed through the frozen backbone's LM head. Let $\mathbf{h}_t$ denote the final normalized hidden state at a scoring position and $\ell_t(v)$ the native logit for vocabulary token v. The Qwen implementation obtains classification logits by selecting the relevant rows of the frozen LM-head matrix. It does not add a trainable classification head.

Table 9: MIntRec2.0 intent verbalizers used by the Qwen classifier. Verbalizers are tokenized in their prompt context.

Intent	Verbalizer	Intent	Verbalizer
Acknowledge	acknowledge	Advise	advise
Agree	agree	Apologise	apologise
Arrange	arrange	Ask for help	help
Asking for opinions	opinions	Care	care
Comfort	comfort	Complain	complain
Confirm	confirm	Criticize	criticize
Doubt	doubt	Emphasize	emphasize
Explain	explain	Flaunt	boast
Greet	greet	Inform	inform
Introduce	introduce	Invite	invite
Joke	joke	Leave	leave
Oppose	oppose	Plan	plan
Praise	praise	Prevent	prevent
Refuse	refuse	Taunt	mock
Thank	thank	Warn	warn

Single-label classification. For a class verbalizer token v_c, define

$$s_c = \ell_t(v_c), \qquad p_c = \frac{\exp(s_c)}{\sum_{k=1}^{K} \exp(s_k)}, \qquad \widehat{y} = \arg\max_c p_c. \tag{26}$$

The verbalizers used by the Qwen classifier are checked to be single tokens in their scoring context. Consequently, this implementation requires no multi-token length normalization. MOSI minimizes $\mathcal{L}_{\mathrm{CE}} = -\log p_y$. For MIntRec2.0, the 30 intent names are mapped to the verbalizers in Table 9, and the training loss is

$$\mathcal{L}_{\mathrm{GCE}} = \frac{1 - p_y^{\gamma}}{\gamma}, \qquad \gamma = 0.7. \tag{27}$$

Here γ denotes the GCE exponent, to distinguish it from the attention query $\mathbf{q}$. Evaluation reports the fraction of instances whose predicted class matches the annotation.

Multi-label movie genre classification. MM-IMDb uses the fixed ordering of 23 genres given below. In the Qwen implementation, a fixed answer scaffold containing 23 space-separated zeros defines the scoring positions. The scaffold is identical for all examples and does not contain the ground-truth genre indicators. At the position preceding the k-th binary token, we compute

$$s_k = \ell_{t_k}(\texttt{1}) - \ell_{t_k}(\texttt{0}), \qquad p_k = \sigma(s_k). \tag{28}$$

The binary cross-entropy objective is

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{23} \sum_{k=1}^{23} [y_k \log p_k + (1 - y_k) \log(1 - p_k)]. \tag{29}$$

Prediction applies a common decision threshold to the genre probabilities. For predicted and reference genre sets $\widehat{Y}_i, Y_i$, the reported metrics are

$$\text{Micro-F1} = \frac{2 \sum_i |\widehat{Y}_i \cap Y_i|}{\sum_i |\widehat{Y}_i| + \sum_i |Y_i|}, \qquad \text{Example-F1} = \frac{1}{N} \sum_i \frac{2|\widehat{Y}_i \cap Y_i|}{|\widehat{Y}_i| + |Y_i|}. \tag{30}$$

Both are multiplied by 100 in the tables.

Audio-visual question answering. For a tokenized reference answer $(y_1, \ldots, y_L)$, including its termination token, MUSIC-AVQA minimizes

$$\mathcal{L}_{\mathrm{ans}} = -\frac{1}{L} \sum_{u=1}^{L} \log p(y_u \mid x, y_{<u}). \tag{31}$$

Losses are averaged over examples. Qwen, MiniCPM, and Gemma use greedy autoregressive short-answer generation during evaluation, with early termination at EOS. Qwen and MiniCPM allow at most 12 and 16 new tokens, respectively. Generated answers are evaluated using the same normalization and matching rule described below.

Scoring lowercases the prediction and reference, removes <unk>, retains the answer after any </think> delimiter, replaces non-alphanumeric punctuation with spaces, and collapses whitespace. A prediction is correct only when the two nonempty normalized strings are identical. Empty or nonmatching outputs are counted as incorrect.

C.7 Task Prompt Templates

The following templates specify the task instructions in the Qwen scoring implementation. Media inputs are supplied through the backbone's native multimodal message format. The text field contains the available text when available and the supplied completion under MT; it is omitted from available-only inputs when text is missing.

MOSI.

```
Input text: [TEXT]. Given the class set ["Positive",
"Neutral", "Negative"], what is the sentiment? Output one
label only.
```

MIntRec2.0.

```
System: You are a precise multimodal intent classifier.
Input text: [TEXT]
Classify the communicative intent. Output exactly one
intent word. The intent is
```

MM-IMDb.

```
Predict the movie genres. Output exactly 23 binary tokens
in this order: [Drama, Comedy, Romance, Thriller, Crime,
Action, Documentary, Adventure, Horror, Mystery, Sci-Fi,
Fantasy, Family, Biography, War, History, Music, Animation,
Musical, Western, Sport, Short, Film-Noir]. Use 1 when a
genre applies and 0 otherwise. Separate tokens with one
ASCII space. Output only the 23 tokens: no brackets,
commas, labels, or prose.
```

The plot and poster are supplied before the output-format instruction. The fixed binary scaffold specifies the 23 evaluation positions.

MUSIC-AVQA.

```
Question: [QUESTION]
Answer with only a short phrase. Answer:
```

C.8 Optimization and Comparison Settings

Each dataset–backbone–completion-provider–missing-configuration combination has a separately trained Gate. Training does not mix missing-modality configurations. Parameter sharing refers to the positions and sources within each trained model, as described in Section B.1. The explicitly identified transfer experiments reuse a source Gate without target-side training.

Within a dataset, all backbone–provider–missing-configuration combinations follow the same training and checkpoint-selection protocol. The checkpoint with the best validation performance under that dataset's selection criterion is retained for test evaluation. MM-IMDb uses the same output-scoring and evaluation protocol across its main experiments and ablations.

Table 10: Reference configuration for Gate training.

Setting	Value
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	0
Gradient clipping norm	1.0
Effective batch size	8
Standard maximum epochs	12
Initial coefficient	$r_0 = 0.5$
Attenuation parameter	$\lambda = \log 10000$ (fixed)
Trainable module	Shared Gate only
Hardware	NVIDIA A800 GPUs

Table 10 summarizes the reference optimization configuration in Section 4.1. The backbone is frozen throughout training. Validation-based selection chooses the Gate checkpoint for subsequent evaluation; test examples are used for reporting task performance. The granularity and input ablations are trained separately with the same training budget as Full Gate.

Baseline, direct completion, and +PE do not optimize downstream model parameters. CARES uses task supervision to train the Gate. The parameter-matched control in Section D.3 uses the same training data, task loss, training budget, and validation-based checkpoint selection protocol as its CARES comparison.

MissRAG and NExT-GPT denote direct completion without CARES. Within each comparison between direct completion and CARES, the completed content and task formulation are retained. +PE adds textual guidance identifying completed modalities as approximate inputs and does not enable or train the Gate.

Source-labeling prompts. PE is appended to an existing system instruction, or introduced as a system message when none is present. Available modalities are marked `Present` and completed modalities are marked `Prototype`. For example, MA uses

```
Input Modality Status: Audio: Prototype; Video: Present;
Input text: Present.

The assistant receives an approximate audio data and uses it
to generate a response as accurately as possible.
```

For MV and MT, the approximation sentence names `visual data` and `input text data`, respectively. For MAV, MAT, and MVT it names `visual and audio data`, `audio and input text data`, and `visual and input text data`, using the plural `uses them`. MUSIC-AVQA keeps the task question present and marks only the missing audio or video as `Prototype`.

The MM-IMDb MT prompt is

```
Input Modality Status: Image: Present; Textual description
of the image: Prototype.

The assistant receives an approximate description of the
image and uses it to generate a response as accurately as
possible.
```

For MI, the status is `Image: Prototype; Textual description of the image: Present.`, and the approximation sentence is `The assistant receives an approximate image and uses it to generate a response as accurately as possible.`

Table 11: Coefficient granularity in the Qwen ablations. Indices i, l, h, t, m denote sample, layer, head, query, and completed modality, respectively. Every row is trained separately.

Variant	Coefficient	Shared computational dimensions
Modality	r_m	Sample, layer, head, query
Sample	$r_{i,m}$	Layer, head, query
Layer	$r_{i,l,m}$	Head, query
Head	$r_{i,l,h,m}$	Query
Full Gate	$r_{i,l,h,t,m}$	None

D Detailed Ablation Settings

D.1 Regulation Granularity

The granularity ablations in Table 2 are trained as separate models. Modality learns an input-independent coefficient for each completed modality. Sample conditions regulation on the sample while sharing its coefficient across computational positions. Layer permits variation between layers, Head additionally permits variation between heads, and Full Gate also permits variation between query positions.

For Layer and Head, aggregation is performed on projected Gate features before the 16-unit hidden layer. Let $\mathbf{u}_{i,l,h,t,m} \in \mathbb{R}^{25}$ be the concatenated feature and $v_{i,l,h,t,m}$ its valid-source indicator. For Layer,

$$\overline{\mathbf{u}}_{i,l,m} = \frac{\sum_{h,t} v_{i,l,h,t,m} \mathbf{u}_{i,l,h,t,m}}{\sum_{h,t} v_{i,l,h,t,m}}, \tag{32}$$

and for Head,

$$\overline{\mathbf{u}}_{i,l,h,m} = \frac{\sum_{t} v_{i,l,h,t,m} \mathbf{u}_{i,l,h,t,m}}{\sum_{t} v_{i,l,h,t,m}}. \tag{33}$$

Only valid positions enter each average. The pooled feature is passed through the same hidden and output layers as Full Gate, followed by the same bounded coefficient mapping. The resulting coefficient is broadcast over the pooled dimensions. These state-conditioned variants retain the same Gate architecture and trainable parameter count; they are not obtained by averaging Full Gate coefficients after training.

D.2 Gate Input Ablations

Input ablations retain the Gate architecture and mask the selected feature block. Remove $\mathbf{q}$, Remove $\mathbf{h}_O$, and Remove $\mathbf{h}_C$ set the corresponding eight-dimensional projected feature to zero after its linear projection and SiLU. This prevents the projection bias from contributing a residual feature. Remove Δ sets the scalar log-mass ratio to zero. All remaining inputs and the attention-reweighting rule are unchanged. Each variant is trained independently under the ablation training protocol.

D.3 Parameter-Matched Tuning

Param-Matched Tuning (PMT) learns a feature-scaling vector at the output of the frozen Qwen2.5-Omni decoder. Let $\mathbf{h}_t$ be its final normalized hidden state. PMT computes

$$\mathbf{h}'_t = \mathbf{g} \odot \mathbf{h}_t, \qquad \boldsymbol{\ell}_t = W_{\mathrm{LM}} \mathbf{h}'_t, \tag{34}$$

where $\mathbf{g} \in \mathbb{R}^{3584}$ is the only trainable parameter vector and W_{LM} is frozen. The vector is initialized to all ones, so its initial output matches the original LM head. It is shared across samples and scoring positions within a trained model.

PMT can change the contribution of hidden features to the task logits and is optimized directly with the task loss. It therefore provides a supervised adaptation control with a parameter budget close to CARES: 3,584 parameters versus 3,529. The training data, task loss, budget, and validation-based selection protocol follow the matched comparison described in Section 4.3. The results compare source-level evidence regulation with this parameter-matched form of output-feature adaptation.

E Routing Analysis Protocols

E.1 Aggregation Across Queries, Samples, and Positions

Panels (a)–(c) of Figure 3 use Qwen2.5-Omni-7B with MissRAG in the MOSI MT configuration. All 686 test instances are included. A query contributes to a completed-source statistic only when it is a nonpadding position with valid available-side and completed-source representations.

For sample i and layer–head position (l, h), let $\mathcal{T}_{i,l,h}$ denote its valid queries. The sample-level coefficient is

$$\overline{r}_{i,l,h} = \frac{1}{|\mathcal{T}_{i,l,h}|} \sum_{t \in \mathcal{T}_{i,l,h}} r_{i,l,h,t}. \tag{35}$$

The heatmaps report the mean and sample standard deviation across instances:

$$\mu_{l,h} = \frac{1}{N} \sum_{i} \overline{r}_{i,l,h}, \qquad s_{l,h} = \sqrt{\frac{1}{N-1} \sum_{i} (\overline{r}_{i,l,h} - \mu_{l,h})^2}. \tag{36}$$

Queries are averaged within each sample before samples are averaged, so longer inputs do not automatically receive larger sample weights. The two panels use separate color scales for their respective statistics.

E.2 Source Attention Before and After Regulation

Native and routed completed-source masses are measured at the same attention operation along the Full-Gate forward trajectory. For MOSI MT there is one completed text source. Its masses are $\alpha_{C,i,l,l,h,t}$ and $\widetilde{\alpha}_{C,i,l,l,h,t}$ before and after the current source reweighting, respectively. We average valid queries within each sample, then average across samples. The layer-wise curves in Figure 3(c) give equal weight to each head within a layer. The overall values give equal weight to all 784 layer–head positions.

Under this aggregation, the completed-source mass decreases from 22.79% to 10.94%. The before and after measurements share the same local query, key, and value state; their difference is the effect of the current source-level reweighting.

E.3 Controlled Text-Completion Conditions

Figure 3(d) uses four text-completion conditions for the same 600 MOSI target samples. Exact supplies the target's original transcript. Correct supplies another transcript with the same sentiment. Irrelevant supplies neutral text. Wrong supplies text with the opposite sentiment. The target's available modalities and prediction task are retained across conditions.

The mean coefficients reported in the main text are 0.458 for Exact, 0.390 for Correct, 0.334 for Irrelevant, and 0.237 for Wrong. The ordering is consistent with greater retention of relevant evidence and stronger attenuation of unrelated or contradictory completions.

E.4 Interpreting the Learned Coefficient Distribution

The bounded parameterization provides a continuous range of regulation strengths. Most learned Gate outputs remain in the interior of this range rather than concentrating near its endpoints. This behavior is consistent with graded retention of completed evidence. The layer–head means and across-sample variation in Figure 3(a,b) describe the complementary positional and sample-dependent aspects of this regulation.

F Extended Evaluation and Transfer Protocols

F.1 Larger Hybrid Backbone

Table 3 evaluates Nemotron-3-Nano-Omni-30B-A3B on all six MOSI missing-modality configurations and both MM-IMDb configurations using MissRAG completion. Baseline uses available

Table 12: Source and target conditions for frozen-Gate transfer. All rows use Qwen2.5-Omni-7B and no target-side Gate training.

Transfer	Source data	Source provider	Target data	Target provider
Cross-dataset	MIntRec2.0	NExT-GPT	MOSI	NExT-GPT
Cross-completion	MOSI	MissRAG	MOSI	NExT-GPT

inputs, MissRAG uses the completed inputs without source regulation, and CARES adds the trained shared Gate. The attention-layer integration is described in Section B.5. Task definitions and reported metrics follow Section C.6.

The reported gains are absolute percentage-point differences between CARES and MissRAG for the same metric. Both MM-IMDb metrics are reported separately. This experiment evaluates source regulation in a hybrid decoder while retaining the native Mamba-2 and MoE computations.

F.2 Frozen-Gate Transfer

Both transfer experiments use Qwen2.5-Omni-7B and evaluate on MOSI with NExT-GPT completions. Table 12 specifies the source and target conditions. Each transfer matches the missing-modality configuration between source and target.

For cross-dataset transfer, the Gate learned for intent recognition is applied to the sentiment-classification task. For cross-completion transfer, the Gate learned with retrieved content is applied with generated content. In both cases, transferred parameters remain frozen. The target task uses its own inputs and scoring rule, and the Gate computes coefficients from the target attention states.

NExT-GPT denotes the target MOSI pipeline with NExT-GPT completion and CARES disabled. Baseline uses only the available target modalities. The average in Table 4 is the arithmetic mean over MA, MV, MT, MAV, MAT, and MVT; parenthesized improvements are percentage-point differences from the corresponding NExT-GPT score. This separates the two transfer settings while keeping the target evaluation and completion provider fixed.