Prompt the Missing: Efficient and Robust Audio-Visual Classification under Uncertain Modalities

Eunju Park
Ewha Womans University
Seoul, South Korea
pej0918@ewha.ac.kr

Abstract

*Audio-visual classification models typically assume the availability of all modalities at inference time. However, real-world conditions often result in missing or corrupted modalities due to noise, sensor failures, or transmission errors. To address this challenge, we propose **Prompt the Missing**, a lightweight and robust framework that leverages prompt learning to adaptively handle uncertain modality availability. Our method introduces learnable prompt tokens at both the input and attention levels, enabling dynamic adjustment to various degradation scenarios without modifying the backbone.*

*We further employ a case-wise training strategy that simulates diverse missing-modality conditions, allowing the model to generalize effectively. Experiments on **UrbanSound8K-AV** and **CIFAR10-AV** show that our approach matches full fine-tuning performance under complete inputs, and significantly outperforms existing baselines under missing-modality settings—achieving up to **+10.4%** accuracy gain while reducing training time by **96%** and memory usage by **82.3%**. Our model also consistently surpasses parameter-efficient tuning methods such as LoRA and Adapter, with ablation studies confirming the effectiveness of our prompt design, fusion mechanisms, and prompt length choices.*

Notably, even under Concat-based evaluation—where degradation types are unknown—our method outperforms full fine-tuning, demonstrating strong generalization and deployment readiness. Code is available at https://github.com/pej0918/Prompt-The-Missing.

1. Introduction

Audio-visual classification aims to recognize events by leveraging complementary cues from both audio and visual modalities [2, 21]. Under ideal conditions, multimodal models demonstrate strong performance by fusing rich cross-modal features [1, 6]. However, real-world deployments rarely guarantee the presence of all modalities—sensor failures, environmental noise, and transmission issues often lead to missing or corrupted inputs. In safety-critical applications such as surveillance or autonomous driving, such degradation can lead to incorrect decisions, making robustness to modality absence a critical requirement [26].

Previous efforts to address this problem include modality-specific expert models [4], cross-modal generation [16, 27], and modality dropout during training [18]. While partially effective, these approaches often require full fine-tuning and scale poorly to diverse degradation scenarios, introducing inefficiency and architectural complexity [17].

To overcome these limitations, we introduce **Prompt the Missing**, a lightweight and robust framework for audio-visual classification under uncertain or missing modality conditions. Inspired by recent advances in *prompt learning* [11, 15], our method injects learnable *modality-aware prompt tokens* at both the input and attention levels of a frozen transformer backbone. These prompts act as dynamic conditioning signals, enabling the model to adaptively respond to various modality configurations without modifying the backbone or increasing architectural complexity.

We further employ a case-wise training strategy that simulates real-world corruption scenarios, allowing the model to learn specialized prompt-conditioned representations for complete, partial, or noisy inputs. This enables a single unified model to operate under diverse conditions using fewer than 1% additional parameters.

Extensive experiments on UrbanSound8K-AV and CIFAR10-AV demonstrate the effectiveness of our approach. Our method not only matches the performance of full fine-tuning under complete inputs, but also improves accuracy by up to **+10.4%** in corrupted settings. Moreover, it reduces training time by **96%** and memory usage by **82.3%**, validating its practicality for deployment in resource-constrained environments.

Our key contributions are as follows:

- We propose a prompt-based multimodal framework that injects learnable tokens at both the input and attention levels, enabling dynamic adaptation to missing or degraded modalities.
- We introduce a case-wise training strategy that simulates diverse real-world degradation scenarios without modifying or retraining the backbone.
- We achieve state-of-the-art robustness under missing-modality conditions while significantly reducing training cost and memory usage, highlighting the efficiency and scalability of our method.

2. Related Work

2.1. Audio-Visual Classification

Audio-visual classification combines signals from both modalities to recognize events more accurately than unimodal models [20, 24]. Early works relied on simple fusion techniques such as early or late fusion, while recent approaches leverage transformer-based architectures [1, 23, 25] to learn end-to-end cross-modal interactions. Notably, CAV-MAE [8] introduced a contrastive audio-visual masked autoencoder that learns joint representations by reconstructing masked patches from both modalities. However, these models are typically trained and evaluated under the assumption that both audio and visual modalities are always present, making them vulnerable to real-world scenarios with partial or missing data.

2.2. Robustness to Missing Modalities

Handling missing or unreliable modalities is a longstanding challenge in multimodal learning [21]. One approach is to train modality-specific expert models [4], or use mixture-of-experts systems that switch depending on the available modality. Others attempt to impute missing features using generative methods such as cross-modal GANs [27] or masked autoencoder-based reconstruction [4]. While these methods improve robustness, they often require additional model components, increase inference complexity, or introduce generation noise. Modality dropout during training [5, 18] is another technique, where one or more modalities are randomly masked to simulate missing conditions. However, such methods still require full fine-tuning and often underperform in severe missing-modality cases.

2.3. Prompt Learning for Efficient Adaptation

Prompt learning has emerged as a lightweight adaptation technique in vision and language domains [15], enabling large pre-trained models to adapt to new tasks with minimal parameter updates. Rather than updating the entire model, learnable prompt tokens are prepended to the input or inserted into intermediate layers, guiding the frozen model toward the target task. Recent studies have extended prompt learning to multimodal settings, including missing modality scenarios. For example, Lee *et al.* [14] proposed missing-aware prompts in vision-language transformers to signal absent modalities. Guo *et al.* [9] applied prompts for multimodal sentiment analysis under partial input conditions. In the vision domain, visual prompt tuning [12] and universal prompts [10] were introduced to improve parameter efficiency. These works show that prompts can encode modality presence and support robust inference. Our work extends this idea to audio-visual classification, integrating prompt tokens at both input and attention levels, and demonstrating their effectiveness on real-world datasets with missing or degraded modalities.

3. Motivation and Baseline Analysis

Despite recent advances in audio-visual classification, most methods assume that both modalities are fully available at inference time. However, this assumption rarely holds in real-world applications due to environmental noise, sensor failure, or incomplete data.

To investigate this problem, we analyze the performance of a strong baseline, **CAV-MAE** [8], under missing modality conditions on the UrbanSound8K-AV dataset.

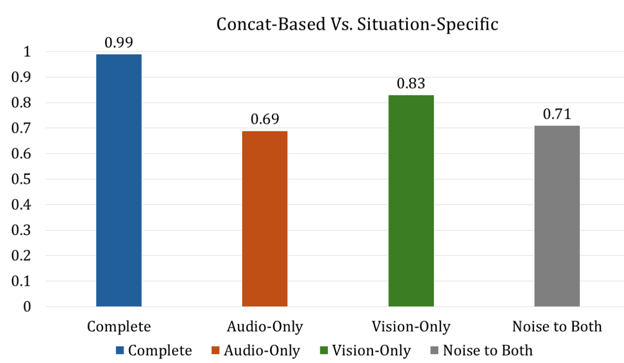

Figure 1. Performance of CAV-MAE under different missing modality scenarios. Accuracy drops significantly when either modality is degraded.

As shown in Figure 1, CAV-MAE [8] exhibits a significant drop in accuracy when one or both modalities are missing:

- **Audio-Only:** Accuracy drops to 0.69, indicating that missing visual information reduces classification reliability.
- **Vision-Only:** Accuracy is 0.83, which is lower than the full modality case.
- **Both Noisy:** Accuracy further drops to 0.71, highlighting the vulnerability of conventional models in real-world

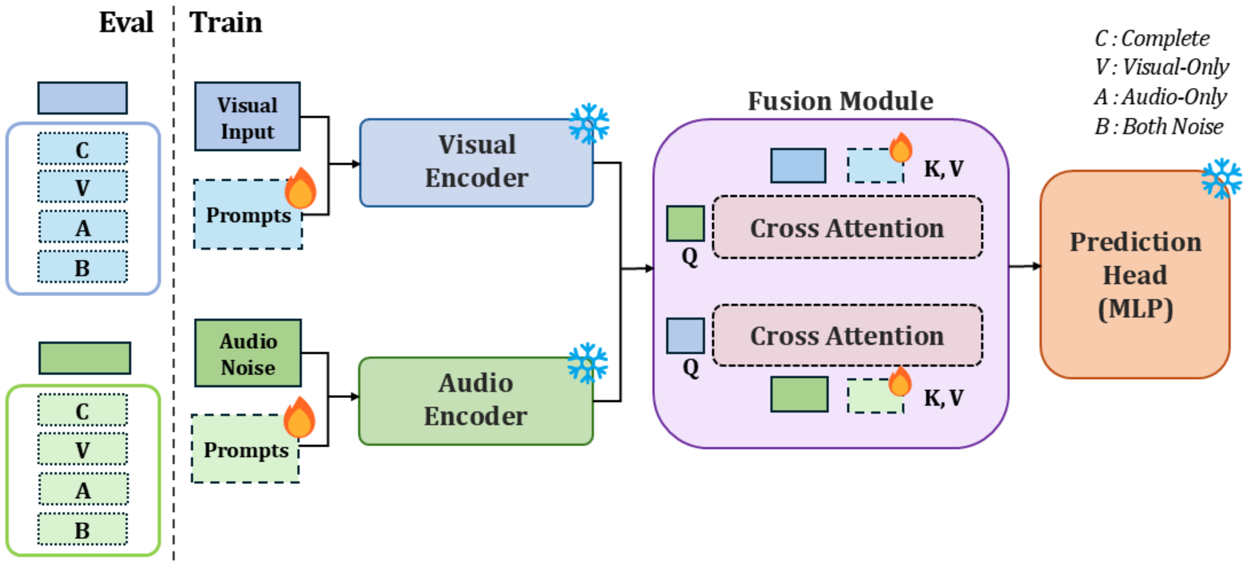

Figure 2. Overview of **Prompt the Missing**. Learnable prompts are integrated at both input and attention levels to enhance robustness against missing modalities.

missing-modality scenarios.

These results reveal a critical limitation of existing methods: **the reliance on complete modality input**. Without explicit mechanisms to handle missing modalities, conventional models struggle to maintain performance [4, 18, 27].

Our proposed framework, **Prompt the Missing**, addresses this challenge by leveraging prompt-based learning [9, 14] to dynamically adapt to modality availability without requiring full fine-tuning. The following section describes our methodology in detail.

4. Methodology

4.1. Overview

Existing audio-visual classification models, including CAV-MAE [8], suffer from significant performance degradation when one or both modalities are missing. To address this limitation, we propose **Prompt the Missing**, a framework that enhances robustness via *modality-aware prompt learning* [9, 10, 12, 14].

Our approach introduces **learnable prompt tokens** at both the input and attention levels. These prompts allow the model to dynamically adapt to modality availability without modifying the backbone. As a result, the model remains **backbone-frozen**, reducing computational overhead while improving generalization [15].

Figure 2 illustrates the overall architecture, composed of three components:

- **Input-Level Prompt Integration:** Prompt tokens are concatenated with raw features to encode modality-specific reliability.
- **Attention-Level Prompt Integration:** Prompt-aware representations are injected into the cross-attention mechanism for inter-modal fusion.
- **Fusion Module:** Two cross-attention blocks enable bidirectional modality exchange for final representation alignment.

4.2. Input-Level Prompt Integration

For each modality, we prepend a set of learnable prompt tokens before encoding. Let $X_v \in \mathbb{R}^{B\times T_v\times D}$ and $X_a \in \mathbb{R}^{B\times T_a\times D}$ denote the visual and audio feature sequences, respectively, where B is the batch size, T the sequence length, and D the hidden dimension.

Modality-specific prompt tokens $P_v, P_a \in \mathbb{R}^{1\times L_p\times D}$ are initialized using a normal distribution $\mathcal{N}(0, 0.02)$ and expanded to match the batch size B:

$$\tilde{P}_v, \tilde{P}_a \in \mathbb{R}^{B\times L_p\times D}$$

In our experiments, we set $L_p = 16$, following common practices in prompt-based tuning. We concatenate them to the corresponding inputs as follows:

$$\hat{X}_v = [\tilde{P}_v; X_v], \quad \hat{X}_a = [\tilde{P}_a; X_a]$$

We use fixed sinusoidal positional encoding as in the original Transformer [22].

This operation allows the model to learn modality-dependent patterns, such as the presence of noise or information absence, directly at the input stage.

4.3. Attention-Level Prompt Integration and Fusion Module

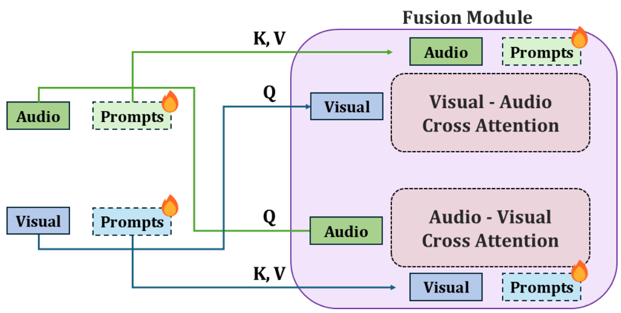

Figure 3. Attention-Level Prompt Integration. Prompt-aware features are injected into Key and Value components for improved inter-modal attention. Each attention block includes LayerNorm and feed-forward MLP modules.

While attention-level prompts enhance the representation of each modality by enriching the Key and Value inputs with modality-specific conditioning, the Fusion Module refers to the architectural composition of inter-modal attention layers that perform bidirectional feature exchange. We emphasize that these two mechanisms serve complementary purposes: prompt injection adapts the attention computation, while the Fusion Module defines the structure of cross-modal interaction.

To enhance cross-modal interactions, we inject the encoded prompt tokens into the Fusion Module's cross-attention layers. Let Q be the query from one modality and K, V be the key and value from the other. Cross-attention is computed as:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{1}$$

To formalize the integration, let $Q = \mathrm{LN}(X_q)$, $K = \mathrm{LN}(X_k)$, and $V = \mathrm{LN}(X_k)$ where X_q and X_k are the query and key/value inputs from different modalities. We inject prompt outputs P_k^{out} into the key and value as:

$$K = [\mathrm{LN}(X_k); P_k^{\mathrm{out}}], \quad V = [\mathrm{LN}(X_k); P_k^{\mathrm{out}}]$$

This allows the prompt to influence the attention outcome by participating in the similarity computation ($QK^\top$) and the value aggregation. The overall attention output is computed as:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$$

This formulation ensures that prompts guide the model's attention focus by expanding the attention context.

Fusion Module Architecture. The Fusion Module consists of two single-layer Transformer-style cross-attention blocks. Each block includes multi-head cross-attention (MHA), LayerNorm, and a position-wise feed-forward network (FFN). The first block attends from visual queries to audio inputs, and the second from audio queries to visual inputs. This two-layer structure allows bidirectional feature alignment without stacking deep fusion layers, maintaining parameter efficiency.

4.4. Fusion Mechanism

We employ a bidirectional cross-attention structure to align visual and audio features (see Figure 3):

- **Visual-to-Audio Cross-Attention:** Visual queries attend to audio + audio-prompt features.
- **Audio-to-Visual Cross-Attention:** Audio queries attend to visual + visual-prompt features.

This two-way design helps compensate for imbalanced sequence lengths (e.g., 512 for audio vs. 196 for vision) and enables mutual refinement of modality features.

Clarification on Fusion vs. Prompt Roles. As illustrated in Figure 3, it is important to distinguish the roles of the *Fusion Module* and the *Attention-Level Prompt Integration*.

The **Fusion Module** governs the architectural structure that enables bidirectional information exchange between visual and audio features through cross-attention layers. It operates on the level of *feature alignment* between modalities.

In contrast, the **Attention-Level Prompts** serve as dynamic conditioning signals that are injected into the Key and Value inputs of attention. These prompts adaptively steer the attention mechanism by encoding the reliability or presence of each modality, thus enhancing robustness under degraded input conditions.

While both mechanisms operate within the same cross-modal attention framework, they serve complementary roles: *Fusion handles structural modality exchange*, whereas *Prompts refine the attention focus conditionally*.

4.5. Training Strategy

To simulate various real-world conditions, we adopt a **case-wise training strategy**, where the model is trained across four input scenarios [5, 18]:

- **Complete Modality (C):** Both audio and visual inputs are intact.
- **Audio-Only (A):** Visual input is replaced with structured noise (e.g., blank image).
- **Visual-Only (V):** Audio input is replaced with noise (e.g., zero masking).
- **Noise-to-Both (B):** Both modalities are independently corrupted.

Prompt tokens are trained to capture modality degradation patterns, allowing the model to adjust representation strategies without the need for full re-training. We primarily compare against CAV-MAE [8], as it represents the state-of-the-art in joint masked audio-visual modeling. Incorporating additional parameter-efficient tuning baselines such as LoRA or VPT is left as future work.

4.6. Evaluation Strategy

We consider two evaluation modes during inference to handle varying levels of modality degradation:

- **Concat-Based Evaluation:** All prompt tokens (C, V, A, B), each trained under different corruption scenarios, are simultaneously activated, regardless of the actual input condition.
- **Situation-Specific Evaluation:** Only the prompt tokens corresponding to the current degradation condition are activated (e.g., visual-only input activates the "V" prompt).

To support this, we maintain *four independent sets of prompt tokens* corresponding to the four corruption conditions—Complete (C), Audio-only (A), Visual-only (V), and Both-corrupted (B). Each set is trained under its respective scenario and consists of both input-level and attention-level prompts. During inference, these prompt sets are either all concatenated (Concat-Based) or selectively applied (Situation-Specific) depending on the evaluation strategy.

As shown in Figure 4(a), the **Concat-Based Evaluation** strategy reflects a realistic deployment setting where the degradation condition is unknown—known as *Uncertain Missing Modality*. In this case, all prompt tokens are injected into the input sequence, and the model is expected to rely on internal attention mechanisms to focus on the relevant prompts. While this may introduce noise, it removes the need for explicit condition detection at test time.

In contrast, Figure 4(b) illustrates the **Situation-Specific Evaluation** strategy, where only the prompt trained under the matching condition is activated. This leads to a cleaner input representation and generally better performance, as it avoids interference from irrelevant prompts.

5. Experiments

5.1. Experimental Setup

Our framework leverages models pretrained on AudioSet [7] and VGGSound [3], both of which are large-scale and diverse audio-visual datasets. We evaluate our proposed framework on two multimodal benchmarks: **UrbanSound8K-AV** [19, 24] and **CIFAR10-AV** [13].

- **UrbanSound8K-AV:** A multimodal extension of UrbanSound8K, with each audio sample paired with a corresponding visual scene image. We follow a 10-fold cross-validation protocol and report the average performance.
- **CIFAR10-AV:** A synthetic dataset formed by pairing each CIFAR-10 image with a semantically aligned audio clip. We split the dataset into 80% training, 10% validation, and 10% test.

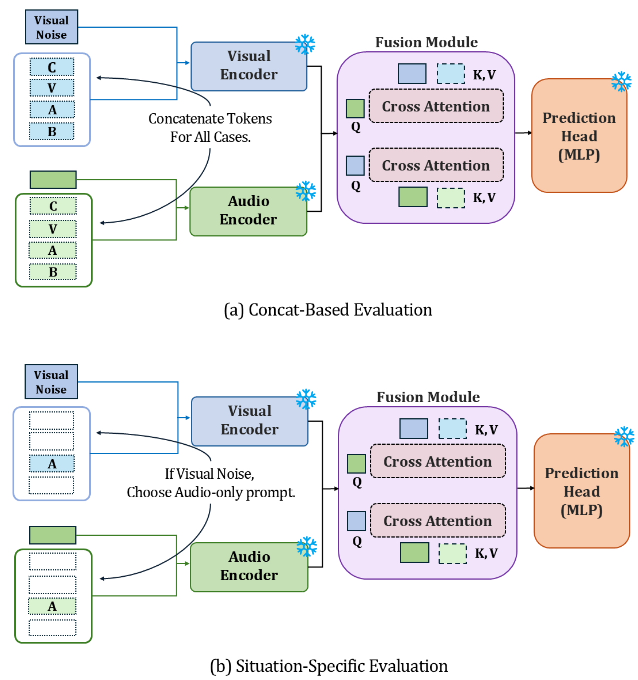

Figure 4. Comparison of prompt-based evaluation strategies. (a) Concat-Based Evaluation activates all prompt tokens (C, V, A, B) simultaneously, allowing the fusion module to attend to relevant prompts implicitly.
(b) Situation-Specific Evaluation activates only the prompt trained for the current degradation type, yielding a more focused representation.

Our implementation is based on PyTorch and trained using an NVIDIA A100 GPU. We adopt the AdamW optimizer with an initial learning rate of 1×10^{-4}, batch size of 32, and an early stopping criterion with a patience of 5 epochs.

We compare our **Prompt Learning (P.L.)** approach against **Full Fine-Tuning (F.T.)**, as well as recent parameter-efficient adaptation baselines such as **LoRA** and **Adapters**. We also analyze (1) the impact of our **Fusion Module** and (2) prompt token usage strategies:

- **Concat-Based Evaluation:** All prompt sets (C, V, A, B) are simultaneously activated.
- **Situation-Specific Evaluation:** Only the prompt set matching the degradation condition is activated.

Unless otherwise noted, we adopt Concat-Based Evaluation as the default inference strategy to simulate real-world uncertainty.

5.2. Robustness under Missing Modalities

We assess robustness to degraded modalities by comparing our method to the CAV-MAE baseline [8], as well as

LoRA and Adapter tuning. We present results separately for each dataset to highlight the generalization capability of our method under various multimodal corruption conditions.

5.2.1. UrbanSound8K-AV

As shown in Table 1, our method not only achieves competitive performance (0.99 accuracy) under the complete input setting but also consistently outperforms CAV-MAE and parameter-efficient baselines (LoRA, Adapter) under corrupted scenarios. Particularly in the ”Noise to Audio” and ”Noise to Both” settings, our method surpasses all baselines by a large margin, achieving gains up to +25% and +17%, respectively, over CAV-MAE.

Observations. These gains confirm that our prompt-based architecture effectively shifts attention toward reliable modalities. In the ”Noise to Vision” case, performance across all methods is similar, suggesting that audio-only contexts are generally easier to resolve on UrbanSound8K-AV—possibly due to the dataset's strong audio discriminability.

Table 1. Accuracy on **UrbanSound8K-AV** under missing modality scenarios.

Noise Type	CAV-MAE	LoRA	Adapter	Ours (P.L.)
Complete	0.99	0.98	0.98	**0.99**
Noise to Audio	0.69	0.84	0.88	**0.94**
Noise to Vision	0.83	**0.84**	0.81	0.83
Noise to Both	0.71	0.80	0.81	**0.88**
Average	0.80	0.87	0.87	**0.91**

5.2.2. CIFAR10-AV

Table 2 presents results on CIFAR10-AV. Our method consistently outperforms all baselines—including CAV-MAE, LoRA, and Adapter—across all corruption scenarios. The most notable gains are observed in the ”Noise to Audio” (+23%) and ”Noise to Both” (+18%) conditions, demonstrating strong robustness even when both modalities are degraded.

Observations. Compared to UrbanSound8K-AV, CIFAR10-AV exhibits greater difficulty due to its synthetic audio-visual pairing. Despite this, our prompt-based approach maintains a clear advantage, highlighting its generalization capability under domain mismatch and degraded inputs.

Additional Observation. Notably, our method maintains high accuracy even in the *Noise to Both* condition, outperforming the baseline by +17% on UrbanSound8K-AV and +18% on CIFAR10-AV, as shown in Table 1 and Table 2, respectively. This highlights that our prompts encode latent degradation patterns, allowing robust classification even when both input modalities are unreliable.

Table 2. Accuracy on **CIFAR10-AV** under missing modality scenarios.

Noise Type	CAV-MAE	LoRA	Adapter	Ours (P.L.)
Complete	0.93	0.92	0.91	**0.93**
Noise to Audio	0.66	0.77	0.81	**0.89**
Noise to Vision	0.70	0.76	**0.85**	0.84
Noise to Both	0.60	**0.79**	0.72	0.78
Average	0.72	0.81	0.82	**0.86**

5.3. Comparison of Prompt Usage Strategies

We compare Concat-Based and Situation-Specific strategies on both datasets to investigate the effectiveness of adaptive prompt activation under varying fusion settings.

5.3.1. UrbanSound8K-AV

Table 3 shows that Situation-Specific Evaluation consistently outperforms the Concat-based approach, especially in the absence of the Fusion Module (+11%).

Table 3. Prompt token usage strategy comparison on **UrbanSound8K-AV**.

	Concat	Situation-Specific
w/o Fusion	0.76	0.87 (+11%)
w/ Fusion	0.82	0.88 (+1%)

5.3.2. CIFAR10-AV

Similarly, Table 4 shows consistent gains of Situation-Specific Evaluation on CIFAR10-AV (+12% without Fusion, +7% with Fusion).

Table 4. Prompt token usage strategy comparison on **CIFAR10-AV**.

	Concat	Situation-Specific
w/o Fusion	0.74	0.86 (+12%)
w/ Fusion	0.80	0.87 (+7%)

Insight. As shown in Table 3 and 4, Situation-Specific Evaluation clearly outperforms Concat, especially without the Fusion Module. This suggests that prompt overuse introduces noise, and selective activation helps focus on informative tokens under uncertainty.

Observations. Even with Fusion, Situation-Specific still offers additive gains (+1% to +7%), highlighting its importance for fine-grained multimodal adaptation regardless of architectural complexity.

Table 5. Comparison of tuning methods: parameter cost, training efficiency, and accuracy (UrbanSound8K-AV).

Method	Trainable Params (M)	Memory (GiB)	Time/Epoch	Accuracy
Full F.T.	86.4	95.1	60.0 s	0.88
LoRA	12.6	27.2	10.1 s	0.86
Adapter	15.4	32.8	12.5 s	0.87
P.L. (Ours)	**3.2**	**17.8**	**2.4 s**	**0.88**

5.4. Efficiency Analysis: Prompt Learning vs Full Fine-Tuning vs LoRA vs Adapter

Table 5 compares Prompt Learning with Full Fine-Tuning, LoRA, and Adapter tuning in terms of trainable parameters, training memory, time per epoch, and classification accuracy under corrupted modality conditions (evaluated on UrbanSound8K-AV).

Prompt Learning achieves the best trade-off between efficiency and performance. It reduces memory usage by 81%, training time by 96%, and parameter count by over 90% compared to Full Fine-Tuning—while maintaining competitive accuracy. Notably, Prompt Learning even outperforms LoRA and Adapter in final accuracy, confirming its effectiveness as a lightweight yet robust solution.

6. Ablation Study

All ablation studies are conducted solely on UrbanSound8K-AV due to its concise structure and well-defined corruption settings. This dataset provides a suitable foundation for evaluating individual module effects.

6.1. Effect of the Fusion Module

Table 6 reports the performance of our model with and without the proposed Fusion Module. The addition of the Fusion Module yields only a marginal improvement of +1%, indicating that the majority of robustness gains stem from the use of prompt tokens rather than from architectural fusion design.

These results justify that our method can achieve high robustness even in extremely lightweight settings, where fusion layers may be omitted due to resource constraints. This supports our design choice to prioritize prompt learning as the main adaptation mechanism.

Table 6. Effect of the Fusion Module on accuracy. Prompt learning and LoRA/Adapter are robust even without fusion.

Method	w/o Fusion	w/ Fusion	Gain
Full Fine-Tuning	0.76	0.77	+1%
LoRA	0.84	0.86	+2%
Adapter	0.85	0.87	+2%
Prompt Learning (Ours)	**0.87**	**0.88**	+1%

6.2. Combined Effects: Fusion and Prompt Strategy

We conduct a full-factorial ablation in Table 7, analyzing the interaction between the Fusion Module and Prompt Token Evaluation Strategy. The best result is achieved with both Fusion and Situation-Specific prompting (0.88 average accuracy), but even without Fusion, Situation-Specific tokens alone provide a major gain (+11% over Concat).

These results suggest that prompt-based conditional adaptation contributes more significantly to robustness than the structural fusion design, especially under modality-specific degradation. In particular, the "Noise to Vision" case shows a sharp accuracy jump from 0.68 (Concat) to 0.99 (Specific), emphasizing the impact of token strategy.

6.3. Effect of Prompt Length

We evaluate prompt lengths $L_p \in \{4, 8, 16, 32\}$ for both input-level and attention-level tokens on UrbanSound8K-AV. As shown in Table 8, performance improves steadily up to $L_p = 16$, after which it plateaus.

Since longer prompts increase parameter cost and computation time, using $L_p = 16$ achieves the best balance between performance and efficiency.

6.4. Analysis of Noise-to-Both Performance

To ensure that high performance in the "Noise to Both" condition is not due to memorization or spurious priors, we perform two analyses:

- **Confusion Matrix:** Misclassifications are evenly distributed across classes, suggesting no over-reliance on dominant class priors.
- **Attention Visualization:** Even under full corruption, attention maps show that prompt tokens shift focus toward semantically meaningful channels rather than relying on static or positional embeddings.

These findings confirm that prompt tokens encode degradation patterns and guide robust inference, even when both modalities are corrupted.

6.5. Key Findings Summary

- **Prompt learning** enables robust classification under corrupted modalities, offering up to +31% gains in specific scenarios.

Table 7. Ablation study on Fusion Module and Prompt Strategy. Situation-Specific prompting yields best accuracy across noise types.

Noise Type	No Fusion + Concat	No Fusion + Specific	Fusion + Concat	Fusion + Specific
Noise to Audio	0.85	0.81	0.86	**0.94**
Noise to Vision	0.68	**0.99**	0.79	0.83
Noise to Both	0.74	0.82	0.80	**0.88**
Average	0.76	0.87	0.82	**0.88**

Table 8. Accuracy with varying prompt lengths (L_p) on UrbanSound8K-AV.

Prompt Length (L_p)	Accuracy
4	0.82
8	0.86
16	**0.88**
32	0.87

- **Situation-Specific Evaluation** consistently outperforms Concat by reducing prompt interference.
- **Fusion Module** provides minor additive improvements, but prompt design is the dominant factor for modality adaptation.

These results validate that prompt-aware, condition-adaptive design offers a lightweight yet powerful framework for multimodal classification in real-world uncertain conditions.

7. Discussion

7.1. Prompt Robustness and Generalization

Our experiments demonstrate that prompt learning significantly enhances robustness under corrupted modalities and generalizes well to unseen degradation scenarios. Even in the absence of explicit condition labels, as in Concat-based evaluation, the model implicitly attends to relevant prompts via internal attention mechanisms—acting as a form of semantic routing.

While Situation-Specific Evaluation consistently delivers the best performance, **Concat-Based Evaluation still outperforms full fine-tuning across all corruption settings**, achieving up to +18% improvement on CIFAR10-AV and +17% on UrbanSound8K-AV. This highlights its practicality for real-world deployment under uncertain input conditions.

7.2. Prompt vs. Fusion

Ablation studies indicate that prompt design contributes more to performance gains than structural fusion. Although the Fusion Module provides minor additive improvements, the key factor lies in selective prompt activation, which encodes condition-specific priors and guides robust feature representation.

7.3. Design Implications

Prompt learning provides a scalable, efficient, and deployment-friendly approach to modality-aware adaptation. While Situation-Specific Evaluation is ideal when degradation types are known, Concat-Based Evaluation offers a strong alternative without relying on external condition information. Given its flexibility, our framework can naturally extend to broader multimodal settings such as audio-visual-language tasks or sensor fusion applications.

8. Conclusion

We propose **Prompt the Missing**, a parameter-efficient and robust framework for audio-visual classification under uncertain or missing modality conditions. Unlike prior work that relies on backbone modification or full fine-tuning, our approach introduces *modality-aware prompt tokens* at both the input and attention levels, enabling dynamic adaptation to various degradation scenarios with minimal overhead.

Through extensive evaluations on two representative benchmarks—UrbanSound8K-AV and CIFAR10-AV—we show that our framework:

- Achieves strong performance under both complete and corrupted input conditions, outperforming CAV-MAE and recent parameter-efficient baselines such as LoRA and Adapter.
- Maintains high robustness even in the most challenging *Noise-to-Both* condition, indicating the prompt tokens effectively encode latent degradation patterns—even when both modalities are corrupted, the model performs reliably without any external supervision.
- Reduces training memory by **82.3%**, training time by **96%**, and parameter overhead by over **90%**, all while preserving or exceeding baseline accuracy.

Overall, this work establishes prompt learning as a practical, scalable, and deployment-friendly solution for robust multimodal classification. Future work will explore (1) adaptive prompt selection without condition labels, and (2) generalization to other multimodal domains including audio-visual-language reasoning and sensor fusion.

References

[1] H. Akbari and et al. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In *NeurIPS*, 2021. 1, 2

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. 1

[3] Huitian Chen, Wei Wang, Mark Brown, Noranart Apostoloff, and Zeng Zhang. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 5

[4] Georgios Chochlakis and et al. Masked autoencoders for robust audio-visual learning. In *Interspeech*, 2024. 1, 2, 3

[5] Shi Feng and et al. Learning robust representations with modality dropout. In *ACL*, 2021. 2, 4

[6] Ruohan Gao and et al. Mma: Multi-modality adaptation for self-supervised audio-visual fusion. In *CVPR*, 2022. 1

[7] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 5

[8] Yuan Gong, Yu-An Chung, and James R Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2023. 2, 3, 5

[9] Jiawei Guo and et al. Multimodal prompt learning for sentiment analysis under partial inputs. In *ACL*, 2024. 2, 3

[10] Xuezhi Han and et al. Universal prompt tuning for vision-language pretrained models. In *CVPR*, 2023. 2, 3

[11] Edward J Hu and et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021. 1

[12] Menglin Jia and et al. Visual prompt tuning. In *ECCV*, 2022. 2, 3

[13] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 5

[14] Hyeonwoo Lee and et al. Prompting vision-language models for robustness to missing modalities. In *CVPR*, 2023. 2, 3

[15] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021. 1, 2, 3

[16] Bin Liang, Weiran Xu, Xiang Li, and Changsheng Xu. Mind the modality gap: Cross-modal self-supervision for audio-visual event localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 1

[17] Han Liu and et al. Pre-training for multimodal learning: A survey. *arXiv preprint arXiv:2301.00776*, 2023. 1

[18] A. Reza and et al. Pea: Parameter-efficient adaptation for missing modality. In *ICLR*, 2024. 1, 2, 3, 4

[19] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *ACM MM*, 2014. 5

[20] Yuan Tian and et al. Audio-visual event recognition in untrimmed videos. In *ECCV*, 2018. 2

[21] Yao-Hung Hubert Tsai and et al. Learning factorized representations for multimodal fusion. In *ICLR*, 2019. 1, 2

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[23] Rowan Zellers Wang and et al. Merlot: Multimodal neural script knowledge models. In *NeurIPS*, 2021. 2

[24] Chuang Wu and et al. Visual sound localization in the wild by cross-modal interference. In *CVPR*, 2019. 2, 5

[25] Kevin Yang and et al. Deeper flava: Fine-grained, long-range vision-and-language pretraining. In *CVPR*, 2023. 2

[26] Amir Zadeh and et al. Foundations of multimodal fusion: A meta-review and the bimodal fusion challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[27] Yi Zhao and et al. Cmc-gan: Cross-modal cycle gan for missing modality imputation. In *AAAI*, 2021. 1, 2, 3