DECORATE YOUR INCOMPLETE MULTIMODAL INPUTS: TOWARDS TEST-TIME ADAPTATION FOR MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

The real-world deployed systems often encounter the incomplete multimodal instances, which incentivize the study of Modality-Missing Learning (MML). Nevertheless, existing methods rely on customized model architectures and training strategies to handle incomplete multimodal inputs. As a result, they are less compatible with the real-world applications, where downstream models are often already designed and trained, and the training process is also often inaccessible due to the privacy issues. To address these limitations, we propose a more practical research problem, **MML-TTA**, which aims to improve the robustness of multimodal models, previously designed and pre-trained on modality-complete samples, to incomplete instances at test time. Accordingly, we introduce a new framework **IM-DECOR**. It treats latent incomplete multimodal inputs as a model-agnostic *adaptation anchor*, and *decorates these inputs to mimic complete ones* observed during training, bridging the distribution shift in MML-TTA caused by input-level structural mismatches. Specifically, IM-DECOR introduces a modality-wise Global Decorator (G-DECOR) and an instance-aware Local Decorator (L-DECOR). They progressively transform incomplete inputs from mimicking *modality-level* distributions of complete data to resembling *instance-specific* complete inputs, enabling robust, on-the-fly adaptation of diverse downstream models to test-time missing inputs. Experiments on six benchmarks under MML-TTA show the efficacy of our method, achieving an average relative improvement of 9.86% over the strongest baseline.

1 INTRODUCTION

Real-world multi-source data (e.g., text, images, audio) have driven the rapid development of the multimodal learning community (Yuan et al., 2025), which integrates these data for impressive performance (Huang et al., 2021). However, the common assumption of deployment-time complete multimodal data often does not hold in the real world due to sensor failures, limited working conditions, or privacy constraints (Chen et al., 2024). As a result, extensive research has been galvanized toward Modality-Missing Learning (MML) (Ma et al., 2021; Wen et al., 2026).

Existing MML methods design *specialized* modality-missing model architectures, such as reconstruction networks (Wang et al., 2023c; Wu et al., 2024), modality-shared networks (Zhuang et al., 2025), along with *dedicated* training pipelines to handle incomplete inputs. Despite their effectiveness, such a **customized architecture design and training for missing modalities** paradigm is less practical for a wide range of multimodal applications (Wang et al., 2023a; Li et al., 2025) (Figure 1). On one hand, downstream-specific models (e.g., clinical models (Jiang et al., 2026), autonomous driving systems (Wen et al., 2026)) are often already deliberately designed, extensively trained, and deployed, making redesign and retraining costly in both development effort and computational resources, and potentially compromising the performance (Lin et al., 2025). On the other hand, privacy constraints on training procedures further limit the feasibility of such tailored solutions (Wang et al., 2021). Furthermore, missing modalities often occur at inference time (Ramazanova et al., 2025), which makes it difficult to anticipate missing patterns and reduces the efficacy of preparatory training strategies, especially as the number of modalities grows (Lin et al., 2025; Jang et al., 2024).

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.

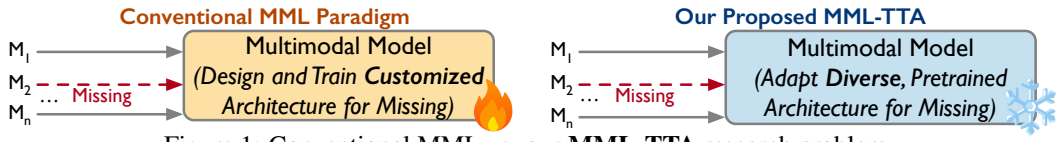


Figure 1: Conventional MML vs. our **MML-TTA** research problem.

Motivated by the success of Test-Time Adaptation (TTA) (Ramazanova et al., 2025), which adapts a pre-trained model from a source domain to an unlabeled, distribution-shifted target domain without access to the training process, we launch a more practical research problem for MML (Figure 1):

MML-TTA: *Downstream models are assumed to be* ***previously designed and fully trained*** *on complete multimodal data, with no access to their original training data and process. The objective of MML-TTA is to adapt these off-the-shelf models on-the-fly* ***at test time*** *to modality-missing inputs.*

To address this open setting, we observe that the distribution shift in MML-TTA arises from the *structural mismatch of multimodal inputs* between training and inference: the fully observed modality sets during training change to partially observed ones encountered at test time, which significantly alters the input composition. As a result, downstream models exposed only to complete training instances struggle to generalize to incomplete ones at deployment (Ma et al., 2022). Moreover, to ensure versatility across broad downstream models, we **wrap** the pre-trained model and treat the latent incomplete multimodal inputs as a **general adaptation anchor**, through which the modality-missing adaptation can be performed without making strong assumptions on specific model architectures.

We propose **IM-DECOR**, a model-agnostic framework [1] that **DECOR**ates **I**ncomplete **M**ultimodal inputs to facilitate test-time adaptation to missing modalities for *diverse* downstream models, *without modifying* their parameters or architectures. IM-DECOR incorporates a modality-wise Global Decorator (**G-DECOR**) and an instance-aware Local Decorator (**L-DECOR**), which *work in tandem to decorate latent incomplete multimodal inputs at test time to mimic complete ones*, boosting the predictions on these samples from downstream models directly at development. Specifically, G-DECOR begins by fetching modality-specific information from the test-time data stream and initializing lightweight modality-level decoration containers. Through an online container update mechanism over the incoming test data stream, G-DECOR continually refines the containers to yield stable global modality proxies for missing modalities. Subsequently, considering the characteristic variations across incomplete samples (Hu et al., 2024; Wang et al., 2023c), L-DECOR transforms the remaining modalities of each sample into learnable queries, which then retrieve the most relevant information from the decoration container to dynamically perform instance-specific local decoration. When *relatively complete samples that contain more than one modality* are available, IM-DECOR additionally introduces an **intra-modal decoration regularization** that encourages the local decorations to align with the ground-truth representations, improving the fidelity of the decorations. Finally, the decorated multimodal inputs are fed into the pre-trained downstream models to foster robust, on-the-fly modality-missing adaptation at test time, while making *zero modifications* on these models to maintain their original pre-trained capabilities. To sum up, our main contributions are as follows:

- We propose a general and practical research problem (**MML-TTA**), which shifts the MML focus, from redesigning and training customized missing-modality models for each downstream task, to *directly adapting* diverse off-the-shelf downstream models to incomplete inputs at deployment.
- We introduce a novel model-agnostic framework, **IM-DECOR**, which treats latent missing-modality inputs as a *general adaptation anchor* across downstream models, and performs global and local decorations on these inputs, realizing robust test-time adaptation for missing modalities.
- Experiments on six benchmarks, covering dual- and tri-modality compositions, diverse backbones, and various test-time missing conditions, show the efficacy of our method. Moreover, IM-DECOR *bypasses the backbone* during backpropagation and updates only the decoration parameters, incurring modest overhead. Code is at `https://anonymous.4open.science/r/IM-DECOR`.

2 Related Work

Modality-Missing Learning. Existing Modality-Missing Learning (MML) methods can be broadly divided into two categories: (1) *Reconstruction-based methods* recover missing modalities through

[1] Agnostic refers to independence from model-specific redesign or retraining, rather than black-box access.

various generative models and strategies (Wu et al., 2024; He et al., 2026a; Du et al., 2026; Wang et al., 2023c;b). Several studies design cross-modal reconstruction networks, such as VAE (Shen et al., 2021) or diffusion models (Wang et al., 2023c), to reconstruct missing modalities using the remaining modalities, while recent studies also employ MLLMs to guide the imputation (Ke et al., 2025). (2) *Reconstruction-free methods* design sophisticated model architectures and training objectives to enhance robustness to missing modalities (Wang et al., 2023a; Wu et al., 2024). A line of work endeavors to align the representations of incomplete samples with those of complete ones via tailored frameworks and training strategies, such as teacher–student distillation (Zhuang et al., 2025; Li et al., 2024; Fang et al., 2026), while others adopt modality-shared architectures with joint learning objectives (Wen et al., 2026). Nevertheless, (i) existing methods heavily rely on model redesign and retraining, undermining their flexibility to be integrated into real-world scenarios, where domain-specific models are often already well-designed and optimized (Wang et al., 2023a). Moreover, (ii) missing patterns are difficult to predict during training, making such preparatory training less effective (Lin et al., 2025). A recent work, MiDl (Ramazanova et al., 2025), explored alleviating the modality-missing problem at test time in downstream task of *egocentric video understanding*. In contrast, we propose a *more general* MML-TTA setup and problem without restricting to *a specific task or model architecture*, and propose IM-DECOR, which is compatible with a wide range of downstream models for on-the-fly modality-missing adaptation at test time.

Test-Time Adaptation. Test-Time Adaptation (TTA) is an efficient and general domain adaptation paradigm, where a source domain pre-trained model adapts to the distribution-shifted, unlabeled target domain, without accessing the (source) training data or process (Liang et al., 2024). Mainstream TTA methods (e.g., Tent (Wang et al., 2021), EATA (Niu et al., 2022)) adopt unsupervised entropy minimization to optimize source models, implicitly encouraging *source-like representations* on shifted target instances to improve the prediction confidence and accuracy (Liang et al., 2020). Recent studies extend TTA to multimodal scenarios (Yang et al., 2024; He et al., 2026b), analyzing unimodal shifts (Chen et al., 2025), mixed multimodal shifts (Guo & Jin, 2025), and concurrent modality shifts (Wang et al., 2025). SuMi (Guo & Jin, 2025) treats modality missing only as a strong-OOD case in wild TTA, without formulating or systematically studying modality-missing TTA as a standalone problem. Borrowing the *source-domain-free* and *target-domain-aware* properties of TTA, we now systematically bridge TTA and MML and formulate MML-TTA as a dedicated research problem.

3 Methodology: Decorate Your Incomplete Multimodal Inputs

Definition of MML. For clarity, we consider a dual-modality scenario with two modalities $m^{(1)}$ and $m^{(2)}$. Each data instance is denoted as (X, Y), where $X = \{x^{(1)}, x^{(2)}\}$ represents the raw modality inputs and Y is the corresponding label. Conventional MML methods design specialized model architectures with dedicated training objectives (Wu et al., 2024; He et al., 2026a), and train the models under artificially simulated missing-modality conditions (e.g., masking modalities):

$$\min_{\theta} \ \mathbb{E}_{(X,Y)\sim\mathcal{D}_{\text{train}}} \left[L_{\text{miss}}\big(F_\theta^s(X), Y\big)\right], \tag{1}$$

where $\mathcal{D}_{\text{train}}$ often contains both complete and masked incomplete multimodal samples (i.e., $\{x^{(1)}, x^{(2)}\}$, $\{x^{(1)}, \varnothing\}$, and $\{x^{(2)}, \varnothing\}$), $F_\theta^s(\cdot)$ denotes the customized modality-missing models, and $L_{\text{miss}}(\cdot)$ represents the training objective that jointly optimizes the task performance and the resilience of $F_\theta^s(\cdot)$ to incomplete multimodal inputs. However, in real-world scenarios, downstream multimodal models (e.g., medical diagnosis models (Jiang et al., 2026)) are often already well-designed and fully trained (Ramazanova et al., 2025), making such *redesign and retraining for missing-modality paradigms* less compatible with these off-the-shelf models and thus less practical.

Definition of MML-TTA. To resolve the above limitations, we formalize a new and practical research problem, MML-TTA. Specifically, we consider a downstream multimodal model F_θ^a that *has been trained* on complete multimodal data $(X, Y) \sim \mathcal{D}_{\text{full}}$, where each $X = \{x^{(1)}, x^{(2)}\}$. For a unified formulation, we decompose F_θ^a as: $F_\theta^a = \{f_\theta^{(1)}, f_\theta^{(2)}, H_\theta, P_\theta\}$, *without* imposing strong assumptions on its architecture or training process, where $f_\theta^{(1)}, f_\theta^{(2)}$ are modality-specific encoders, H_θ is the modality fusion network, and P_θ is the classifier. At test time, following the practical online TTA setting (Wang et al., 2021), *unlabeled* multimodal data continually arrive in a sequence of batches $\{\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_n, \cdots\}$. Due to real-world factors such as sensor failures (Wang et al., 2023a), these

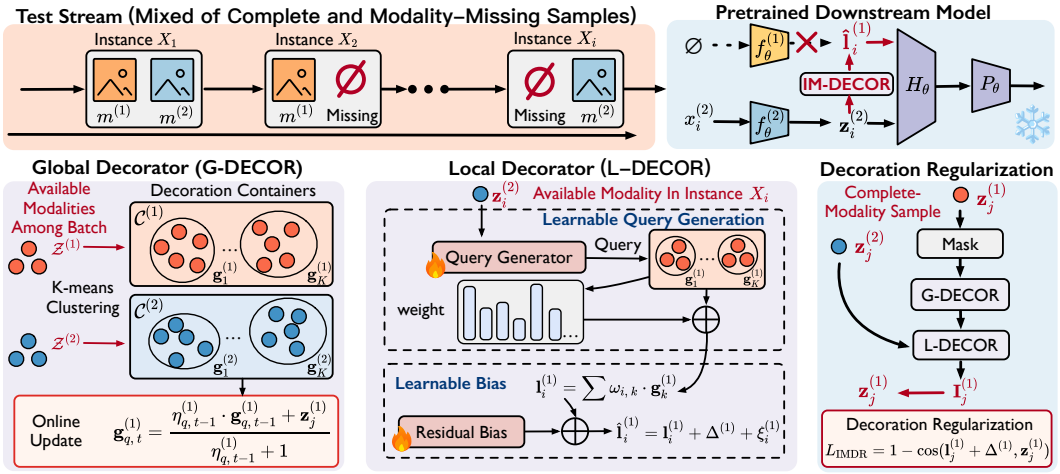


Figure 2: Overview of IM-DECOR, which decorates your incomplete inputs for robust MML-TTA.

deployment-time data follow a different distribution $\mathcal{D}_{\text{miss}}$ [2], where each sample $X \sim \mathcal{D}_{\text{miss}}$ may be either complete or partially observed (modality-missing), i.e., $X = \{x^{(1)}, x^{(2)}\}$, $\{x^{(1)}, \varnothing\}$, or $\{x^{(2)}, \varnothing\}$. MML-TTA aims to directly adapt F_θ^a (potentially update its parameters) on each incoming batch $\mathcal{B}_i$ in an online manner, improving performance on each batch immediately.

Workflow of IM-DECOR. Since the distribution shift between $\mathcal{D}_{\text{full}}$ and $\mathcal{D}_{\text{miss}}$ mainly stems from the *structural mismatch of the instances*, and to ensure compatibility with diverse F_θ^a, IM-DECOR identifies latent incomplete multimodal inputs as a *general adaptation anchor* and *decorates them to mimic complete ones*, bridging the domain gap for improved performance while keeping the pre-trained F_θ^a *frozen* to prevent catastrophic forgetting (Figure 2). Specifically, IM-DECOR first performs **global decoration** by maintaining modality-wise decoration containers (Sec. 3.1). Subsequently, it conducts instance-level **local decoration**, activating the remaining modalities in each sample as learnable queries to aggregate relevant information from the containers (Sec. 3.2). When relatively complete instances are available, the locally decorated inputs are further aligned with the ground-truth representations (Sec. 3.3). Finally, the decorated inputs are fed into F_θ^a for adaptation (Sec. 3.3).

3.1 Modality-Wise Global Decoration

To effectively decorate incomplete multimodal instances to resemble complete ones, we propose a modality-wise Global Decorator (G-DECOR). It first initializes lightweight decoration containers for each modality in the early batch, and continually refines the containers in an online manner as more test data arrive, yielding stable and robust *global modality proxies* for the missing modalities.

Decoration Container Initialization. Specifically, taking the initialization of the decoration container for modality $m^{(1)}$ as an example, G-DECOR first creates an empty container $\mathcal{C}^{(1)} = \emptyset$. Subsequently, G-DECOR yields the set $\mathcal{Z}^{(1)}$ of modality-specific representations from samples in the incoming batch that contain modality $m^{(1)}$ via a forward pass from the corresponding encoder $f_\theta^{(1)}$:

$$\mathcal{Z}^{(1)} = \{\mathbf{z}_1^{(1)}, \mathbf{z}_2^{(1)}, \cdots, \mathbf{z}_N^{(1)}\}, \quad \mathbf{z}_i^{(1)} = f_\theta^{(1)}(x_i^{(1)}), \tag{2}$$

where $\mathbf{z}_i^{(1)} \in \mathbb{R}^{d^{(1)}}$ [3] is the modality-specific representation of sample X_i, $x_i^{(1)}$ is the modality data of X_i, N is the number of samples that hold modality $m^{(1)}$ in the batch. To ensure lightweight, compact yet representative modality proxies, we further apply K-means clustering (Bottou & Bengio, 1994) on $\mathcal{Z}^{(1)}$, yielding a set of K cluster prototypes that initialize the decoration container $\mathcal{C}^{(1)} = \left\{\mathbf{g}_k^{(1)}\right\}_{k=1}^{K}$:

$$\arg\min_{\mathcal{S}_k} \sum_{k=1}^{K} \sum_{\mathbf{z}_i^{(1)} \in \mathcal{S}_k} \left\| \mathbf{z}_i^{(1)} - \mathbf{g}_k^{(1)} \right\|^2, \quad \mathbf{g}_k^{(1)} = \frac{1}{|\mathcal{S}_k|} \sum_{\mathbf{z}_i^{(1)} \in \mathcal{S}_k} \mathbf{z}_i^{(1)}, \tag{3}$$

[2] Any modality may be missing in individual samples. However, the data stream collectively retains all modalities, thereby preventing the original task from degrading into a modality-reduced one.

[3] We present modality representation without sequence dimension, with sequential formulation in App. A.

where $\{\mathcal{S}_k\}_{k=1}^{K}$ denotes the partition of modality-specific representations into K clusters ($K \ll N$), and $\mathbf{g}_k^{(1)} \in \mathbb{R}^{d^{(1)}}$ represents the prototype (centroid) of the k-th cluster.

Online Decoration Container Update. At time step t, where $\mathcal{C}_{t-1}^{(1)}$ is non-empty, we further leverage samples from the incoming batch $\mathcal{B}_t$ to update and refine the container. Since we only maintain the lightweight cluster prototypes in $\mathcal{C}_{t-1}^{(1)}$, we apply an online K-means update mechanism (Bottou & Bengio, 1994), progressively refining the container in an online manner. For sample $X_j \in \mathcal{B}_t$ that holds the modality $m^{(1)}$, we assign its representation $\mathbf{z}_j^{(1)}$ to the nearest prototype in $\mathcal{C}_{t-1}^{(1)}$ and update:

$$q = \arg\min_{k \in \{1,\dots,K\}} \|\mathbf{z}_j^{(1)} - \mathbf{g}_{k,\,t-1}^{(1)}\|^2, \quad \mathbf{g}_{q,\,t}^{(1)} = \frac{\eta_{q,\,t-1}^{(1)} \cdot \mathbf{g}_{q,\,t-1}^{(1)} + \mathbf{z}_j^{(1)}}{\eta_{q,\,t-1}^{(1)} + 1}, \tag{4}$$

where $\mathbf{g}_{q,\,t-1}^{(1)}$ is the nearest prototype for X_j, and $\eta_{q,\,t-1}^{(1)}$ is the counter of the q-th prototype indicating the number of previously assigned samples. Through continual refinement over the incoming test data flow, the G-DECOR gradually forms stable modality proxies that approximate the global distribution.

3.2 Instance-Aware Local Decoration

After deriving the modality-wise global decorations, a naive approach is to directly allocate them to modality-missing instances for the test-time missing adaptation. Considering the varying characteristics across samples (Wang et al., 2023c), we further propose a new instance-aware Local Decorator (L-DECOR). It leverages the remaining modalities as conditions to refine the specific decorations for each instance. Specifically, taking the decoration of $X_i = \{x_i^{(2)}, \varnothing\}$ that loses the modality $m^{(1)}$ as an example, L-DECOR transforms its remaining modality into a learnable query to retrieve and aggregate the relevant information from the decoration container [4], yielding the local decoration:

$$\mathbf{q}_i^{(2)} = \phi_{m^{(2)} \to m^{(1)}}(f_\theta^{(2)}(x_i^{(2)})), \tag{5}$$

$$\mathbf{l}_i^{(1)} = \sum_{\mathbf{g}_k^{(1)} \in \mathcal{C}^{(1)}} \omega_{i,\,k} \cdot \mathbf{g}_k^{(1)}, \quad \omega_{i,\,k} = \frac{\exp\Big(\mathrm{sim}\Big(\mathbf{q}_i^{(2)},\ \mathbf{g}_k^{(1)}\Big)/\tau\Big)}{\sum_{\mathbf{g}_r^{(1)} \in \mathcal{C}^{(1)}} \exp\Big(\mathrm{sim}\Big(\mathbf{q}_i^{(2)},\ \mathbf{g}_r^{(1)}\Big)/\tau\Big)}, \tag{6}$$

where $\phi_{m^{(2)} \to m^{(1)}}$ is the projection network (implemented with MLP) that converts the remaining modality $x_i^{(2)}$ into the learnable query $\mathbf{q}_i^{(2)} \in \mathbb{R}^{d^{(1)}}$, $\mathrm{sim}(,)$ calculates the cosine similarities between the $\mathbf{q}_i^{(2)}$ and each $\mathbf{g}_k^{(1)}$ in the container, and $\mathbf{l}_i^{(1)}$ is the local decoration for the X_i. To improve the expressive capacity of $\mathbf{l}_i^{(1)}$ beyond the retrieved prototypes, we add small learnable residual biases:

$$\hat{\mathbf{l}}_i^{(1)} = \mathbf{l}_i^{(1)} + \Delta^{(1)} + \xi_i^{(1)}, \tag{7}$$

where $\Delta^{(1)} \in \mathbb{R}^{d^{(1)}}$ and $\xi_i^{(1)} \in \mathbb{R}^{d^{(1)}}$ denote sample-shared and sample-specific biases for X_i, respectively, and $\hat{\mathbf{l}}_i^{(1)}$ represents the final local decoration for sample X_i.

3.3 Intra-Modal Decoration Regularization & Modality-Missing Adaptation

Since the real-world environment may also include relatively complete samples with *more than one available modality* (Lin et al., 2025), we propose a new Intra-Modal Decoration Regularization, which leverages the latent cross-modal relationships within these samples as additional supervision to further improve the quality of local decoration. Specifically, under the dual-modality case, we mask each modality in the *complete samples* within $\mathcal{B}$ and apply the regularization on its local decoration:

$$L_{\text{IMDR}} = \frac{1}{|\mathcal{B}^{\text{full}}|} \sum_{X_j \in \mathcal{B}^{\text{full}}} \Big[(1 - \cos(\mathbf{l}_j^{(1)} + \Delta^{(1)}, \mathbf{z}_j^{(1)})) + (1 - \cos(\mathbf{l}_j^{(2)} + \Delta^{(2)}, \mathbf{z}_j^{(2)}))\Big], \tag{8}$$

where $\mathcal{B}^{\text{full}}$ denotes the set of modality-complete samples within $\mathcal{B}$. Notably, we do not apply the regularization to $\hat{\mathbf{l}}_j^{(m)}$, as it includes sample-wise $\xi_j^{(m)}$ that do not generalize to incomplete samples.

[4] We omit the superscript of time step t on the decoration container $\mathcal{C}^{(1)}$, prototype $\mathbf{g}_k^{(1)}$, and batch $\mathcal{B}$.

Algorithm 1 Algorithm of IM-DECOR (Dual-modality samples with $m^{(1)}$ decoration as an example).

Input: A pre-trained downstream model $F_\theta^a = \{f_\theta^{(1)}, f_\theta^{(2)}, H_\theta, P_\theta\}$; test stream $\{\mathcal{B}_t\}_{t=1}^T$.
Output: Predictions $\hat{y}_i$ for samples in each $\mathcal{B}_t$.

1: Initialize global modality-wise decoration container $\mathcal{C}^{(1)} = \emptyset$.
2: **for** each incoming batch $\mathcal{B}_t$ **do**
3: Extract representations $\mathbf{z}^{(1)}$ from the samples in $\mathcal{B}_t$ that have modality $m^{(1)}$ by forwarding them though $f_\theta^{(1)}$, and update $\mathcal{C}^{(1)}$ via (online) K-means strategy. ▷ Global Decoration
4: For each sample $X_i \in \mathcal{B}_t$ that misses $m^{(1)}$, transform the modality $x_i^{(2)}$ into a learnable query $\mathbf{q}_i^{(2)}$, and retrieve and aggregate the local decoration $\hat{\mathbf{l}}_i^{(1)}$ from $\mathcal{C}^{(1)}$. ▷ Local Decoration
5: Calculate L_{IMDR} if complete samples exist in $\mathcal{B}_t$. ▷ Intra-Modal Decoration Regularization
6: Update the decoration parameters by minimizing L_{EM} (and L_{IMDR} when available).
7: Perform prediction: $\hat{y}_i = P_\theta(H_\theta(\hat{\mathbf{l}}_i^{(1)}, f_\theta^{(2)}(x_i^{(2)})))$.
8: **end for**

Modality-Missing Adaptation. Finally, we integrate both the decorated modality input and the remaining available modalities in each incomplete sample to perform the prediction. Specifically, taking the sample X_i with only modality $m^{(2)}$ available as an example, the prediction is given by:

$$\hat{y}_i = P_\theta(H_\theta(\hat{\mathbf{l}}_i^{(1)}, f_\theta^{(2)}(x_i^{(2)}))). \tag{9}$$

During the test time, we also adopt the classical unsupervised entropy minimization (Liang et al., 2020) objective on samples in the incoming batch $\mathcal{B}$ to optimize the decoration parameters:

$$L_{\text{EM}} = \frac{1}{|\mathcal{B}|} \sum_{X_i \in \mathcal{B}} \mathcal{H}\big(P_\theta(\cdot \mid X_i)\big) = -\frac{1}{|\mathcal{B}|} \sum_{X_i \in \mathcal{B}} \sum_{c=1}^{C} \hat{y}_{i,c} \log \hat{y}_{i,c}, \tag{10}$$

where C is the total number of classes. As a result, the complete optimization objective is the combination of L_{EM} and L_{IMDR} (when the relatively complete samples are available). Nevertheless, IM-DECOR does not rely on complete samples, and remains effective even when no such instances exist at test time (cf. Figure 4). Notably, the F_θ^a remains *frozen* during the adaptation to prevent catastrophic forgetting. For complete samples, the decorations are employed only for additional optimization, while *inference is performed on the original modalities without decoration*, preventing F_θ^a from degrading on such in-domain samples during long-term adaptation. To mitigate parameter drift during adaptation (Wang et al., 2022), we reset the sample-specific learnable biases after each adaptation batch. The algorithm of IM-DECOR is provided in Algorithm 1. Since the G-DECOR is designed to be modality-wise, it can scale to multiple modalities, and we provide the mathematical derivation of IM-DECOR under multiple modalities in App. B. The complexity and efficiency analysis are provided in App. C, and the mathematical analysis is in App. D. The behavior of IM-DECOR under boundary test-time modality-missing scenarios is discussed in App. E.

4 Experiments

Benchmark. To showcase the downstream-agnostic feature of IM-DECOR, we evaluate it on six multimodal benchmarks across diverse tasks, domains, and modality compositions. Specifically, for the dual-modality setting, we select ❶ **MM-IMDb** (Arevalo et al., 2017): multi-label movie genre classification with text and image modalities, evaluated using F1-Macro, ❷ **Food101** (Wang et al., 2015): 101-category food classification with text and image modalities, evaluated using Accuracy, ❸ **Kinetics50** (Carreira & Zisserman, 2017): 50-category action classification with video and audio modalities, evaluated using Accuracy, and ❹ **VGGSound** (Chen et al., 2020): sound event recognition with video and audio modalities, evaluated using Accuracy. For the tri-modality setting, we select two multimodal sentiment classification benchmarks: ❺ **IEMOCAP** (Busso et al., 2008) and ❻ **MOSI** (Zadeh et al., 2016), which involve text, audio, and video modalities. We follow (Tsai et al., 2019) and adopt Accuracy and F1 for these two benchmarks.

Source Model. Without strong assumptions on the downstream model structures, we adopt a general multimodal architecture for the source model (Radford et al., 2021; Yang et al., 2024), consisting of

Table 1: Performance (%) of baselines and IM-DECOR on dual-modality benchmarks under a 70% missing rate across diverse missing types. Com. denotes the combination of two missing types, where each is missing at a rate of 35%. The best results are in **bold** and the second are underlined.

		MM-IMDb				Food101				Kinetics50				VGGSound			
Methods	**Venue**	**T**	**I**	**Com.**	**Avg.**	**T**	**I**	**Com.**	**Avg.**	**A**	**V**	**Com.**	**Avg.**	**A**	**V**	**Com.**	**Avg.**
Source	N/A	39.28	31.79	37.12	36.06	76.69	83.89	80.38	80.32	70.36	55.96	62.33	62.88	34.44	54.07	44.50	44.34
Tent	ICLR'21	34.76	31.76	34.56	33.69	77.17	83.42	78.13	79.57	69.67	54.10	61.39	61.72	20.37	53.00	36.54	36.64
EATA	ICML'22	38.39	30.85	36.41	35.22	72.27	84.26	79.06	78.53	69.55	56.20	62.61	62.79	32.59	54.58	43.79	43.65
READ	ICLR'24	34.28	31.35	35.81	33.81	77.52	85.22	80.09	80.94	69.38	56.41	62.45	62.75	30.54	47.59	37.84	38.66
SuMi	ICLR'25	38.86	30.64	36.42	35.31	76.40	81.87	79.07	79.11	69.63	56.24	62.90	62.92	15.11	34.75	23.95	24.60
MiDl	ICLR'25	39.48	31.76	37.74	36.33	77.02	84.72	80.97	80.90	70.80	56.69	63.02	63.50	33.83	54.41	44.17	44.14
PTA	NeurIPS'25	35.66	29.31	36.63	33.87	66.21	74.08	72.21	70.83	69.43	55.66	60.81	61.97	25.32	43.54	38.13	35.66
DASP	CVPR'26	41.09	34.22	38.20	37.84	77.62	84.69	80.56	80.96	61.39	52.72	56.45	56.85	29.72	48.94	38.03	38.90
IM-DECOR	N/A	**48.22**	**48.88**	**48.99**	**48.70**	**79.32**	**87.76**	**83.76**	**83.61**	**73.20**	**58.19**	**64.72**	**65.37**	**37.63**	**55.24**	**46.55**	**46.47**

Table 2: Performance (%) of baselines and IM-DECOR on tri-modality benchmarks under a 70% missing rate across diverse missing types. The best results are in **bold** and the second are underlined.

IEMOCAP		**A**		**T**		**V**		**A-T**		**A-V**		**T-V**		**Avg.**	
Methods	**Venue**	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Source	N/A	56.40	55.81	47.76	44.21	56.93	56.21	47.12	40.56	55.12	53.96	44.46	40.84	51.30	48.60
Tent	ICLR'21	55.76	54.16	41.90	37.81	55.86	54.92	46.27	40.69	54.16	51.48	45.10	41.54	49.84	46.77
EATA	ICML'22	56.72	55.63	44.56	40.83	55.65	54.88	46.27	41.03	55.22	53.18	47.12	43.26	50.92	48.14
READ	ICLR'24	50.85	50.83	46.38	44.41	51.07	51.08	43.18	41.67	50.32	50.69	44.46	42.36	47.71	46.84
SuMi	ICLR'25	55.22	53.97	46.80	42.79	56.82	55.96	46.91	42.05	55.33	53.82	48.61	44.09	51.62	48.78
MiDl	ICLR'25	54.90	53.94	46.27	42.67	54.37	53.83	46.48	41.90	54.26	52.75	47.97	44.08	50.71	48.20
PTA	NeurIPS'25	47.97	48.60	39.66	39.97	49.04	49.54	36.46	37.44	46.16	45.89	33.05	33.46	42.06	42.48
DASP	CVPR'26	38.06	35.23	20.90	17.94	36.89	33.26	23.24	19.31	36.46	32.40	21.43	17.89	29.50	26.01
IM-DECOR	N/A	**57.68**	**57.27**	**52.56**	**49.36**	**58.10**	**57.44**	**49.36**	**45.97**	**56.40**	**55.77**	**53.20**	**48.30**	**54.55**	**52.35**
MOSI		**A**		**T**		**V**		**A-T**		**A-V**		**T-V**		**Avg.**	
Methods	**Venue**	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Source	N/A	36.01	25.59	21.14	13.72	32.65	21.11	20.41	12.79	31.49	20.10	19.24	11.84	26.82	17.53
Tent	ICLR'21	31.34	21.65	19.97	11.86	29.30	18.48	20.55	12.65	27.99	17.47	19.97	07.72	24.85	14.97
EATA	ICML'22	31.34	22.05	20.26	12.47	30.61	19.90	20.55	12.83	29.88	19.22	20.55	08.69	25.53	15.86
READ	ICLR'24	34.99	24.09	21.43	13.62	35.28	22.73	19.97	12.36	31.20	20.86	20.99	12.90	27.31	17.76
SuMi	ICLR'25	35.86	26.08	19.68	12.44	31.92	21.46	20.12	12.47	31.20	20.29	20.85	13.67	26.61	17.74
MiDl	ICLR'25	33.09	23.37	19.97	12.79	32.22	21.28	19.83	12.05	29.15	18.69	21.87	13.49	26.02	16.95
PTA	NeurIPS'25	35.13	21.53	22.45	13.47	35.57	21.31	22.01	12.43	35.42	21.64	20.85	11.98	28.57	17.06
DASP	CVPR'26	27.99	11.47	20.41	08.92	28.28	11.56	20.12	08.91	26.53	10.10	20.55	11.31	23.98	10.38
IM-DECOR	N/A	**37.76**	**26.89**	**24.05**	**16.11**	**37.17**	**26.33**	**22.59**	**15.35**	**39.21**	**27.63**	**22.45**	**14.76**	**30.54**	**21.18**

modality-specific encoders, a modality fusion network, and a classification network. After training on the corresponding training set of each dataset (without any modality-missing training), we select the best-performing model as the final source model. The detailed implementation is in App. F.2.

Evaluation Protocol of MML-TTA. Under MML-TTA, the source model is already designed and pre-trained on modality-complete data. At test time, both modality-missing and complete samples arrive sequentially in batches. Aligning with MML (Hu et al., 2024; Lee et al., 2023), we simulate the batch-wise modality missing according to the type τ and rates η. For instance, $\tau =$ T and $\eta = 70\%$ indicates that, in a given batch, 70% of samples miss the text, while $\tau =$ A-T indicates 70% of samples miss both audio and text. In main evaluation, we keep τ and η consistent across incoming batches, and also evaluate scenarios with dynamic types and rates across batches in Sec. 4.2.

Baselines. Since the MML-TTA setting combines MML and TTA, while most MML methods require redesigning and retraining the source model, we therefore adopt 7 competitive TTA methods as baselines: Tent (Wang et al., 2021), EATA (Niu et al., 2022), READ (Yang et al., 2024), SuMi (Guo & Jin, 2025), MiDl (Ramazanova et al., 2025), PTA (Wang et al., 2025), and DASP (He et al., 2026b). The baseline implementation details under MML-TTA setup are provided in App. F.3.

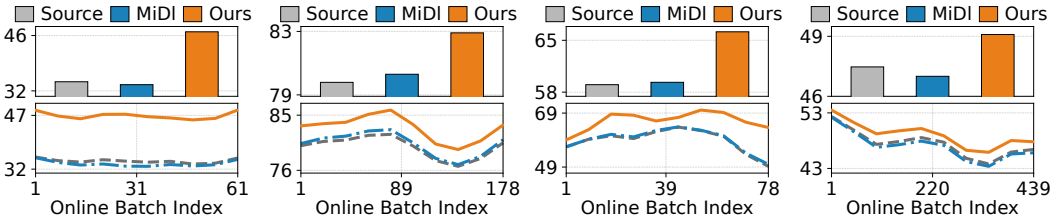
Implementation Details. All experiments use a single NVIDIA RTX 4090 GPU. We use AdamW optimizer with a learning rate of 1×10^{-4}, weight decay of 10^{-2}, and batch size of 128. The source model remains frozen during adaptation, and only the decoration parameters are optimized.

4.1 Main Performance under MML-TTA

From Table 1 and Table 2, we observe that most TTA methods exhibit limited or even worse performance than the source model, as they are primarily designed for synthetic distribution shifts

Table 3: Ablation study (%) under 70% missing rate on four datasets.

Module	Variant	MM-IMDb			Food101			Kinetics50			VGGSound		
		T	I	Com.	T	I	Com.	A	V	Com.	A	V	Com.
G-DECOR	w/o Online	46.62	46.88	46.71	78.17	85.72	83.67	72.11	56.38	63.63	36.14	54.99	45.16
	w/o G-DECOR	40.60	38.23	39.96	77.12	84.14	82.11	71.24	56.11	62.29	34.79	53.07	44.25
L-DECOR	w/o Residual Bias	47.67	47.35	47.62	78.28	85.78	83.72	72.28	57.16	64.03	36.61	55.02	46.06
	w/o Regularization	48.14	47.96	47.99	79.28	86.77	83.75	72.28	58.08	64.07	37.61	55.03	46.09
	w/o L-DECOR	45.10	46.04	45.99	77.26	84.76	82.75	71.95	56.47	62.87	35.65	54.23	45.09
IM-DECOR	All	**48.22**	**48.88**	**48.99**	**79.32**	**87.76**	**83.76**	**73.20**	**58.19**	**64.72**	**37.63**	**55.24**	**46.55**



(a) MM-IMDb. (b) Food101. (c) Kinetics50. (d) VGGSound.

Figure 3: Performance of IM-DECOR and baseline under continually changing missing patterns.

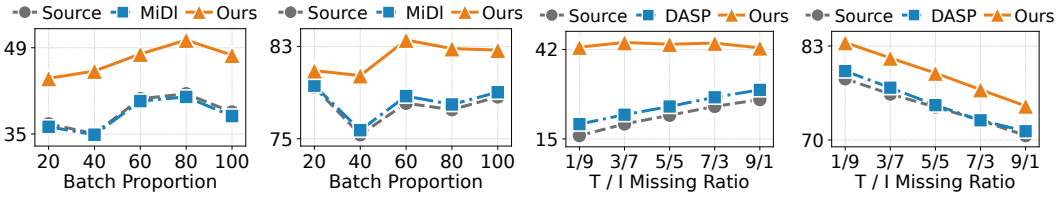
(e.g., adding snow), where the representations of shifted (noisy) instances can be implicitly aligned with the source distribution via adapting normalization statistics with unsupervised objectives (e.g., entropy minimization) (Wang et al., 2021; Niu et al., 2022; Guo & Jin, 2025). However, under MML-TTA, the distribution shift between training and inference arises from a *structural mismatch of multimodal inputs* (certain modalities completely vanish), and is not sufficiently addressed by such implicit methods. Notably, although MiDl attempts to address MML at test time on egocentric videos, it relies on the modality-complete samples for optimization, and its alignment strategies are also indirect for such structural shifts, incurring suboptimal performance. In contrast, IM-DECOR decorates modality-missing samples at test time to mimic complete training ones via a progressively refined process from G-DECOR to L-DECOR, explicitly bridging the domain shifts under MML-TTA and achieving superior performance. Furthermore, IM-DECOR gets rid of updating the source model for missing adaptation like MiDl, avoiding degradation on complete in-domain samples.

4.2 Further Analysis

Ablation Study. To dissect the role of components in IM-DECOR, we conduct ablative analysis, with results in Table 3. For G-DECOR, we design variants: ❶ w/o Online, where the online decoration container update mechanism is removed, and ❷ w/o G-DECOR, where the prototypes in the container are randomly initialized (also without online update). We observe degraded performance in these variants, indicating that the quality of global decoration, which serves as global modality proxies for the missing modalities, is crucial. For L-DECOR, we design three variants: ❸ w/o Residual Bias, where the learnable bias is removed, ❹ w/o Regularization, where L_{IMDR} is removed, and ❺ w/o L-DECOR, where the prototypes within the global decoration container are simply averaged to decorate incomplete inputs. For variant ❸, removing the residual bias limits the expressive capacity of L-DECOR, leading to suboptimal performance. For variant ❹, we observe *a modest performance decline*, indicating that IM-DECOR does not heavily rely on complete samples for adaptation. For variant ❺, applying decoration in an instance-agnostic manner results in performance degradation, as the diverse characteristics of different samples require specific decoration.

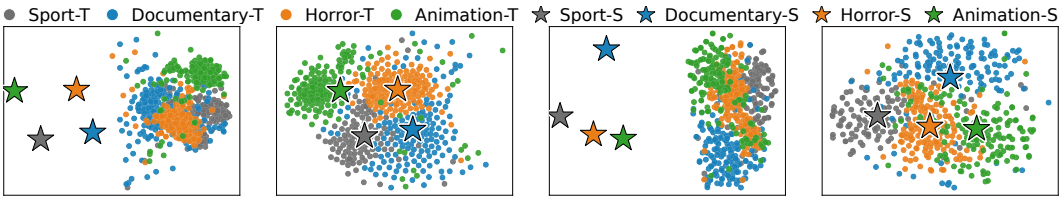
Robustness to Continually Changing Missing Patterns. To validate the robustness of IM-DECOR under dynamic missing patterns encountered at real-world deployment, we conduct experiments on dual-modality benchmarks, where the missing type τ continually changes in each incoming batch (randomly sampled from three types mentioned in Table 1), and rate η remains 70%. As shown in Figure 3, comparing with baseline MiDl, IM-DECOR achieves both superior final performance (the above figures) and relatively stable progression over the online data stream (the below figures), since G-DECOR progressively refines the global decoration container with the continually incoming data, while L-DECOR performs robust decoration for each incomplete sample based on the container.

Robustness to Extreme Missing Scenarios. In real-world deployment, extreme missing patterns often arise, such as a modality being entirely missing over prolonged periods or the complete lack



(a) MM-IMDb. (b) Food101. (c) MM-IMDb. (d) Food101.

Figure 4: Performance of IM-DECOR and baselines under extreme conditions. (a)–(b): Some batches miss certain modality, (c)–(d): No complete samples exist.



(a) MiDl (Text-Only). (b) Ours (Text-Only). (c) MiDl (Image-Only). (d) Ours (Image-Only).

Figure 5: Visualization of samples from source (S) and target (T) domains under MML-TTA.

of modality-complete samples. To evaluate these extreme conditions, we first conduct ❶ **extreme modality-wise missing experiment**, where a certain modality is entirely missing ($\eta = 100\%$) over some contiguous segments of the online test stream. We simulate this on MM-IMDb and Food101 under the text-missing scenario, where the missing ratio η alternates between 70% and 100% across contiguous segments of the online batches, with each segment accounting for 20% of the total batches (stream). In Figure 4(a)–(b), we report the performance of IM-DECOR and MiDl within each segment (i.e., the value at $x = 20$ is performance within $0\% \rightarrow 20\%$ of the online batches.). We observe that, comparing with MiDl, which relies on complete samples for adaptation, IM-DECOR remains robust under this extreme scenario, as G-DECOR maintains lightweight yet accumulative decoration containers that store previously observed decoration resources, enabling reliable decoration even when certain modalities are entirely missing over contiguous periods. Secondly, we conduct ❷ **complete-sample-free experiment**, where no modality-complete samples exist at test time. We simulate this on MM-IMDb and Food101 by conducting five independent evaluations over the entire test stream, where the missing rate ratios of image and text vary from $10\% : 90\%$ to $90\% : 10\%$, and results are in Figure 4(c)–(d). Since MiDl fails to tackle this scenario, we compare IM-DECOR with baseline DASP. We observe a similar conclusion as in the ablation study: even without regularization on complete samples, IM-DECOR can achieve compelling performance under such extreme conditions.

Visualization of the Decoration. To deeply evaluate the efficacy of the decoration from IM-DECOR in bridging distribution gaps for incomplete samples in MML-TTA, we conduct visualizations. Specifically, on MM-IMDb dataset, following (Press et al., 2024) while adjusting our MML-TTA setting, we regard the test set without modality-missing as the source domain, and test set with image- or text-only samples as the target domain. We then visualize the mean representations of four film classes, "Sport", "Documentary", "Horror", and "Animation", from the source data, along with the (decorated) multimodal fused representations from IM-DECOR and MiDl (trained under 70% image or text missing) on 150 sampled (target) image- or text-only instances per class. As shown in Figure 5, comparing with MiDl where representations are dispersed and deviate from the source, the decorated representations from IM-DECOR are compact and closer to their corresponding source classes.

5 Conclusion

In this study, we propose a novel MML-TTA setting that shifts conventional MML research toward adapting diverse downstream multimodal models, pre-trained on complete multimodal data, to incomplete samples at test time. We then propose IM-DECOR, which decorates latent incomplete multimodal inputs to mimic complete ones, explicitly bridging the distribution gap in MML-TTA for effective modality-missing adaptation, while leaving the downstream models unchanged. Extensive experiments on six benchmarks and decoration visualizations show the efficacy of our method.

AI USE STATEMENT

In this work, we used generative AI tools only to aid and polish writing. We did not use generative AI tools to conduct the core empirical analysis or to make final scientific judgments, and other disclosure categories are not applicable to this work. We reviewed all AI-assisted content and take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies test-time adaptation to missing modalities and does not involve new data collection or intervention with human subjects. Our experiments are conducted on established multimodal benchmarks using existing datasets and pretrained models. As a general adaptation framework, IM-DECOR may be applied to deployment scenarios involving sensitive or safety-critical domains. Its improved robustness to missing modalities should not be interpreted as a guarantee of reliability in such applications. Like the underlying datasets and pretrained models, IM-DECOR may inherit existing biases, privacy risks, or representation disparities, and practical deployment should therefore follow appropriate domain-specific validation, privacy protection, and applicable regulations. We are not aware of additional ethical risks specific to the proposed methodology beyond those inherited from the underlying data, models, and application domains.

REPRODUCIBILITY STATEMENT

We provide extensive details to facilitate the reproducibility of our work. The complete formulation and optimization procedure of IM-DECOR are described in Sec. 3 and Algorithm 1, with its extensions to sequential representations and multiple modalities provided in Apps. A and B. Detailed descriptions of all benchmarks, data preprocessing, source-model architectures, baseline implementations, IM-DECOR implementations, and hyperparameters are provided in App. F. Additional experiments across missing rates, model architectures, and ablation settings are reported in App. G. We further provide the assumptions and complete derivations for our theoretical analysis in App. D. Finally, an anonymous code repository containing the implementation and experimental configurations is provided with the submission.

REFERENCES

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Leon Bottou and Yoshua Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 7, 1994.

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

João Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, 2017.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A Large-Scale Audio-Visual Dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Mengxi Chen, Fei Zhang, Zihua Zhao, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Probabilistic conformal distillation for enhancing missing modality robustness. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:36218–36242, 2024.

MingCai Chen, Baoming Zhang, Zongbo Han, Wenyu Jiang, Yanmeng Wang, Shuai Feng, Yuntao Du., and Bingkun BAO. Test-Time Selective Adaptation for Uni-Modal Distribution Shift in Multi-Modal Data. In *International Conference on Machine Learning (ICML)*, 2025.

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.

Xiang Fang, Wanlong Fang, Changshuo Wang, Keke Tang, Daizong Liu, Siyi Wang, and Wei Ji. Towards unified vision-language models with incomplete multi-modal inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 3858–3866, 2026.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27253–27266, 2022a.

Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R Glass. Contrastive audio-visual masked autoencoder. In *International Conference on Learning Representations (ICLR)*, 2022b.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Zirun Guo and Tao Jin. Smoothing the Shift: Towards Stable Test-Time Adaptation under Complex Multimodal Noises. In *International Conference on Learning Representations (ICLR)*, 2025.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1726–1736, 2024.

Zirun Guo, Tao Jin, Wenlong Xu, Wang Lin, and Yangyang Wu. Bridging the Gap for Test-Time Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (16):16987–16995, 2025.

Ningxin He, Yongheng Deng, Sheng Yue, Yongjian Fu, Zhangzehui, and Tiegang Gao. Rag4dmc: Retrieval-Augmented Generation for Data-Level Modality Completion. In *International Conference on Learning Representations (ICLR)*, 2026a.

Yongbo He, Zirun Guo, and Tao Jin. Decoupling stability and plasticity for multi-modal test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep Correlated Prompting for Visual Recognition with Missing Modalities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems (NeurIPS)*, 34:10944–10956, 2021.

Jaehyuk Jang, Yooseung Wang, and Changick Kim. Towards robust multimodal prompting with missing modalities. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8070–8074. IEEE, 2024.

Juntao Jiang, Jiangning Zhang, Yali Bi, Jinsheng Bai, Weixuan Liu, Weiwei Jin, Zhucun Xue, Yong Liu, Xiaobin Hu, and Shuicheng Yan. M3cotbench: Benchmark chain-of-thought of mllms in medical image understanding. In *International Conference on Learning Representations (ICLR)*, 2026.

Guanzhou Ke, Shengfeng He, Xiao Li Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and HeXing Su. Knowledge Bridger: Towards Training-Free Missing Modality Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25864–25873, 2025.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning (ICML)*, pp. 5583–5594. PMLR, 2021.

Marine Le Morvan, Nicolas Prost, Julie Josse, Erwan Scornet, and Gaël Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pp. 3165–3174. PMLR, 2020.

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, 2023.

Jiacheng Li and Songhe Feng. Bridging Modalities via Progressive Re-alignment for Multimodal Test-Time Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12458–12468, 2024.

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, 2025.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In *International Conference on Machine Learning (ICML)*, pp. 6028–6039, 2020.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision (IJCV)*, 133(1):31–64, 2024.

Ronghao Lin, Qiaolin He, Sijie Mai, Ying Zeng, Aolin Xiong, Li Huang, Yap-Peng Tan, and Haifeng Hu. Cyin: Cyclic informative latent space for bridging complete and incomplete multimodal learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 2302–2310, 2021.

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, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient Test-Time Model Adaptation without Forgetting. In *International Conference on Machine Learning (ICML)*, pp. 16888–16905, 2022.

Ori Press, Ravid Shwartz-Ziv, Yann Lecun, and Matthias Bethge. The entropy enigma: Success and failure of entropy minimization. In *International Conference on Machine Learning (ICML)*, pp. 41064–41085. PMLR, 2024.

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

Merey Ramazanova, A. Pardo, Bernard Ghanem, and Motasem Alfarra. Test-Time Adaptation for Combating Missing Modalities in Egocentric Videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Bernhard Scholkopf, Kah-Kay Sung, Christopher JC Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, and Vladimir Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing*, 45(11):2758–2765, 1997.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018.

Meng Shen, Huaizheng Zhang, Yixin Cao, Fan Yang, and Yonggang Wen. Missing data imputation for solar yield prediction using temporal multi-modal variational auto-encoder. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 2558–2566, 2021.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2019.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Guowei Wang, Fan Lyu, and Changxing Ding. Partition-Then-Adapt: Combating Prediction Bias for Reliable Multi-Modal Test-Time Adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15878–15887, 2023a.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7201–7211, 2022.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *IEEE International Conference on Multimedia & Expo Workshops (ICME Workshops)*, pp. 1–6. IEEE, 2015.

Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22025–22034, 2023b.

Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:17117–17128, 2023c.

Ziqiang Wang, Zhixiang Chi, Yanan Wu, Li Gu, Zhi Liu, Konstantinos Plataniotis, and Yang Wang. Distribution alignment for fully test-time adaptation with dynamic online data streams. In *European Conference on Computer Vision (ECCV)*, pp. 332–349. Springer, 2024.

Jiageng Wen, Shengjie Zhao, Bing Li, Jiafeng Huang, Kenan Ye, and Hao Deng. Single-modal-operable multimodal collaborative perception. In *International Conference on Learning Representations (ICLR)*, 2026.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

Wenxin Xu, Hexin Jiang, et al. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2024.

Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the Class Weight Bias: Weighted Maximum Mean Discrepancy for Unsupervised Domain Adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 945–954, 2017.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xiaocui Peng. Test-time Adaptation against Multi-modal Reliability Bias. In *International Conference on Learning Representations (ICLR)*, 2024.

Yuan Yuan, Zhaojian Li, and Bin Zhao. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*, 57(7):1–34, 2025.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-philippe Morency. Mosi: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *arXiv.org*, abs/1606.06259, 2016.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards Robust Multimodal Sentiment Analysis with Incomplete Data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. Cmad: Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4626–4636, 2025.

APPENDIX

A Formulation under Sequential Modality Representation **16**

B Mathematical Derivation under Multiple Modalities **17**

C Complexity and Efficiency Analysis **18**

C.1 Complexity Analysis of IM-DECOR 18

C.2 Efficiency Analysis of IM-DECOR 18

D Mathematical Analysis of IM-DECOR **19**

E Boundary Behaviour of IM-DECOR **20**

F Detailed Experimental Setup **21**

F.1 Benchmarks 21

F.2 Source Model 21

F.3 Baselines 22

F.4 Implementation Details 23

G Additional Experiments **23**

G.1 Main Performance under MML-TTA across Other Missing Rates 23

G.2 Ablation Studies on Tri-modal Benchmarks 23

G.3 Additional Ablation Studies for IM-DECOR 24

G.4 Quantitative Analysis of the Decoration using MMD 24

G.5 Performance on the Early-Fusion Source Model 26

G.6 Comparison with Conventional Modality Missing Methods 27

G.7 Comparison with Unimodal Method 28

G.8 Further Analysis on L-DECOR for Instance-Aware Decoration 28

H Discussion of MML-TTA, Design of IM-DECOR, and Baselines **28**

H.1 Why IM-DECOR is Model-Agnostic and Its Boundary 28

H.2 What Fundamentally Distinguishes IM-DECOR from Existing TTA Methods . . . 29

H.3 Why SuMi and MiDl are Limited under MML-TTA 29

H.4 Why MML-TTA Assumes Stream-Level Modality Coverage 29

I Limitation & Future Work **29**

I.1 Extension of MML-TTA Research Problem 30

I.2 Extensible Design of DECOR with Potential Future Direction 30

J Broader Impacts **30**

A FORMULATION UNDER SEQUENTIAL MODALITY REPRESENTATION

In this section, we provide a more general formulation of IM-DECOR under sequential modality representation. Following the formulation in the main paper, we take the dual-modality scenario as an example, where each modality-complete sample has two modalities: $X_i = \{x_i^{(1)}, x_i^{(2)}\}$.

Modality-Wise Global Decoration. Taking the decoration of modality $m^{(1)}$ as an example, G-DECOR first creates an empty decoration container $\mathcal{C}^{(1)} = \emptyset$. Subsequently, it yields the set $\mathcal{Z}^{(1)}$ of sequential modality-specific representations from the samples in the incoming batch that contain modality $m^{(1)}$ via a forward pass from the corresponding encoder $f_\theta^{(1)}$:

$$\mathcal{Z}^{(1)} = \{\mathbf{z}_1^{(1)}, \mathbf{z}_2^{(1)}, \cdots, \mathbf{z}_N^{(1)}\}, \quad \mathbf{z}_i^{(1)} = f_\theta^{(1)}(x_i^{(1)}), \tag{11}$$

where $\mathbf{z}_i^{(1)} \in \mathbb{R}^{S^{(1)} \times d^{(1)}}$ is the sequential representation of modality $m^{(1)}$ for sample X_i; $S^{(1)}$ is the sequence length and $d^{(1)}$ is the feature dimension. $x_i^{(1)}$ is the modality data of $m^{(1)}$ for X_i, and N is the number of samples that hold modality $m^{(1)}$ in the batch.

Different from the main paper, the decoration container here maintains the sequential prototypes. Nevertheless, to ensure efficiency, we apply K-means clustering (Bottou & Bengio, 1994) on the summary modality-specific representations in the $\mathcal{Z}^{(1)}$: $\bar{\mathbf{z}}_i^{(1)} = \mathrm{Mean}(\mathbf{z}_i^{(1)}) \in \mathbb{R}^{d^{(1)}}$, yielding a set of K cluster prototypes that initialize the sequential decoration container $\mathcal{C}^{(1)} = \left\{\mathbf{g}_k^{(1)}\right\}_{k=1}^{K}$:

$$\arg\min_{\mathcal{S}_k} \sum_{k=1}^{K} \sum_{\mathbf{z}_i^{(1)} \in \mathcal{S}_k} \left\| \bar{\mathbf{z}}_i^{(1)} - \bar{\mathbf{g}}_k^{(1)} \right\|^2, \quad \mathbf{g}_k^{(1)} = \frac{1}{|\mathcal{S}_k|} \sum_{\mathbf{z}_i^{(1)} \in \mathcal{S}_k} \mathbf{z}_i^{(1)}, \quad \bar{\mathbf{g}}_k^{(1)} = \mathrm{Mean}(\mathbf{g}_k^{(1)}), \tag{12}$$

where $\{\mathcal{S}_k\}_{k=1}^{K}$ denotes the partition of sequential modality representations into K clusters ($K \ll N$), $\mathbf{g}_k^{(1)} \in \mathbb{R}^{S^{(1)} \times d^{(1)}}$ is the sequential prototype stored in the decoration container, and $\bar{\mathbf{g}}_k^{(1)} \in \mathbb{R}^{d^{(1)}}$ is its summary representation. At time step t, where $\mathcal{C}_{t-1}^{(1)}$ is non-empty, we further leverage samples from the incoming batch $\mathcal{B}_t$ to update and refine the container via online K-means update:

$$q = \arg\min_{k \in \{1,\ldots,K\}} \|\bar{\mathbf{z}}_j^{(1)} - \bar{\mathbf{g}}_{k,\,t-1}^{(1)}\|^2, \quad \mathbf{g}_{q,\,t}^{(1)} = \frac{\eta_{q,\,t-1}^{(1)} \cdot \mathbf{g}_{q,\,t-1}^{(1)} + \mathbf{z}_j^{(1)}}{\eta_{q,\,t-1}^{(1)} + 1}, \tag{13}$$

where $\bar{\mathbf{g}}_{k,\,t-1}^{(1)} = \mathrm{Mean}(\mathbf{g}_{k,\,t-1}^{(1)})$ is the summary of the k-th sequential prototype, and $\eta_{q,\,t-1}^{(1)}$ is the counter of the q-th prototype indicating the number of previously assigned samples.

Instance-Aware Local Decoration. Taking the decoration of $X_i = \{x_i^{(2)}, \varnothing\}$ as an example, L-DECOR first yields the sequential representation $\mathbf{z}_i^{(2)} = f_\theta^{(2)}(x_i^{(2)})$ of its remaining modality, and then transforms its summary representation $\bar{\mathbf{z}}_i^{(2)} = \mathrm{Mean}(\mathbf{z}_i^{(2)})$ into a learnable query to retrieve and aggregate the semantically relevant information from the container, yielding the local decoration:

$$\mathbf{q}_i^{(2)} = \phi_{m^{(2)} \to m^{(1)}}(\bar{\mathbf{z}}_i^{(2)}), \tag{14}$$

$$\mathbf{l}_i^{(1)} = \sum_{\mathbf{g}_k^{(1)} \in \mathcal{C}^{(1)}} \omega_{i,\,k} \cdot \mathbf{g}_k^{(1)}, \quad \omega_{i,\,k} = \frac{\exp\left(\mathrm{sim}\left(\mathbf{q}_i^{(2)},\ \bar{\mathbf{g}}_k^{(1)}\right)/\tau\right)}{\sum_{\mathbf{g}_k^{(1)} \in \mathcal{C}^{(1)}} \exp\left(\mathrm{sim}\left(\mathbf{q}_i^{(2)},\ \bar{\mathbf{g}}_k^{(1)}\right)/\tau\right)}, \tag{15}$$

where $\phi_{m^{(2)} \to m^{(1)}}$ is the projection network (implemented with MLP) that converts the summary of the remaining modality into the learnable query $\mathbf{q}_i^{(2)} \in \mathbb{R}^{d^{(1)}}$, $\mathrm{sim}(\cdot,\cdot)$ calculates the cosine similarities between $\mathbf{q}_i^{(2)}$ and the summary $\bar{\mathbf{g}}_k^{(1)}$ of each sequential prototype in the container, and $\mathbf{l}_i^{(1)} \in \mathbb{R}^{S^{(1)} \times d^{(1)}}$ is the sequential local decoration for X_i. To improve the expressive capacity of $\mathbf{l}_i^{(1)}$ beyond the retrieved prototypes, we add small learnable residual biases:

$$\hat{\mathbf{l}}_i^{(1)} = \mathbf{l}_i^{(1)} + \Delta^{(1)} + \xi_i^{(1)}, \tag{16}$$

here $\Delta^{(1)} \in \mathbb{R}^{d^{(1)}}$ and $\xi_i^{(1)} \in \mathbb{R}^{d^{(1)}}$ are sample-shared and sample-specific biases for X_i, respectively. To avoid excessive computational overhead, we apply these non-sequential learnable biases via *broadcasted addition*. $\hat{\mathbf{l}}_i^{(1)} \in \mathbb{R}^{S^{(1)} \times d^{(1)}}$ is the final sequential local decoration for X_i. The following formulation follows that of the main paper and is omitted here for brevity.

B MATHEMATICAL DERIVATION UNDER MULTIPLE MODALITIES

In this section, we provide the formulation of IM-DECOR under multiple modalities. Since the G-DECOR is modality-wise, its formulation remains the same as that in the main paper and can be directly applied to each modality independently, and is omitted here. The main differences arise only in the ❶ **instance-aware local decoration** and the ❷ **intra-modal decoration regularization**, where multiple remaining modalities will jointly condition the decoration of each target (missing) modality. For clarity, we omit the sequence dimension of the modality-specific representations in this section.

Instance-Aware Local Decoration. We consider a general scenario, where the full modality set is $\mathcal{M} = \{m^{(1)}, m^{(2)}, \cdots, m^{(|\mathcal{M}|)}\}$. For a given sample X_i, we represent its available modality set as $\mathcal{A}_i \subset \mathcal{M}$ and its missing modality set as $\mathcal{U}_i = \mathcal{M} \backslash \mathcal{A}_i$. Suppose $|\mathcal{A}_i| = A < |\mathcal{M}|$. The L-DECOR treats each remaining modality as a query and aggregates the conditioned decorations from all remaining modalities to form the final instance-aware local decoration. Specifically, for each target modality $m^{(u)} \in \mathcal{U}_i$, a naive extension of the dual-modality scenario is to learn an independent projection network $\phi_{m^{(v)} \to m^{(u)}}$ for every modality pair, resulting in a quadratic growth in the number of projection networks with respect to the number of modalities. To avoid this, we decompose the pair-wise modality projection into a two-stage mapping with an *intermediate dimension* d:

$$\mathbf{q}_i^{(v \to u)} = \phi^{\text{out}}_{m^{(u)}} \left(\phi^{\text{in}}_{m^{(v)}} \left(f_\theta^{(v)}(x_i^{(v)}) \right) \right), \quad m^{(v)} \in \mathcal{A}_i,\ m^{(u)} \in \mathcal{U}_i, \tag{17}$$

where $\phi^{\text{in}}_{m^{(v)}} : \mathbb{R}^{d^{(v)}} \to \mathbb{R}^d$ maps modality $m^{(v)}$ into the shared intermediate space, and $\phi^{\text{out}}_{m^{(u)}} : \mathbb{R}^d \to \mathbb{R}^{d^{(u)}}$ maps the shared representation into the target modality space. In this way, the number of projection mapping is reduced to $2|\mathcal{M}|$, and the mapping is implemented as two lightweight MLP layers for each modality. After obtaining the queries from all available modalities, to avoid the quadratic cost caused by separate modality retrieval, we first aggregate them into a unified query for each target modality, and then retrieve the relevant information from the decoration container $\mathcal{C}^{(u)}$ of the target modality to form the instance-aware local decoration:

$$\mathbf{l}_i^{(u)} = \sum_{\mathbf{g}_k^{(u)} \in \mathcal{C}^{(u)}} \omega_{i,\,k}^{(u)} \cdot \mathbf{g}_k^{(u)}, \quad \bar{\mathbf{q}}_i^{(u)} = \text{Agg}_{m^{(v)} \in \mathcal{A}_i} \left(\mathbf{q}_i^{(v \to u)} \right), \tag{18}$$

$$\omega_{i,\,k}^{(u)} = \frac{\exp\left(\text{sim}\left(\bar{\mathbf{q}}_i^{(u)},\ \mathbf{g}_k^{(u)} \right) / \tau \right)}{\sum_{\mathbf{g}_r^{(u)} \in \mathcal{C}^{(u)}} \exp\left(\text{sim}\left(\bar{\mathbf{q}}_i^{(u)},\ \mathbf{g}_r^{(u)} \right) / \tau \right)}, \tag{19}$$

where $\omega_{i,\,k}^{(u)}$ measures the relevance between the query generated from the available modalities and the k-th prototype in the target container, $\text{Agg}(\cdot)$ represents the aggregation operator that is implemented with average pooling. The final local decoration $\hat{\mathbf{l}}_i^{(u)}$ for modality $m^{(u)}$ still follows the same residual formulation as mentioned in the main paper:

$$\hat{\mathbf{l}}_i^{(u)} = \mathbf{l}_i^{(u)} + \Delta^{(u)} + \xi_i^{(u)}. \tag{20}$$

Intra-Modal Decoration Regularization. Different from the dual-modality setting, we do not enumerate all possible modality subsets as conditions for supervision of the intra-modal decoration regularization, which would incur unaffordable complexity. Instead, for any sample X_i that contains A available modalities with $A > 1$ (we name it as "relatively complete sample"), we perform one alignment training for each available modality by treating it as the target modality and using *all the other available modalities* as the joint conditions. Specifically, for each $m^{(u)} \in \mathcal{A}_i$, we construct its local decoration using the condition set $\mathcal{A}_i \backslash \{m^{(u)}\}$ with the additive formulation above, and align it with the ground-truth representation $\mathbf{z}_i^{(u)}$. The resulting regularization in each online batch is:

$$L^{\text{multi}}_{\text{IMDR}} = \frac{1}{|\mathcal{B}^{>1}|} \sum_{X_i \in \mathcal{B}^{>1}} \frac{1}{|\mathcal{A}_i|} \sum_{m^{(u)} \in \mathcal{A}_i} \left(1 - \cos(\mathbf{l}_i^{(u)} + \Delta^{(u)}, \mathbf{z}_i^{(u)}) \right), \tag{21}$$

where $\mathcal{B}^{>1}$ denotes the set of samples in the batch that contain more than one available modality. Under this implementation, the condition decoration of each remaining modality for the target modality is optimized, while maintaining minimal overhead.

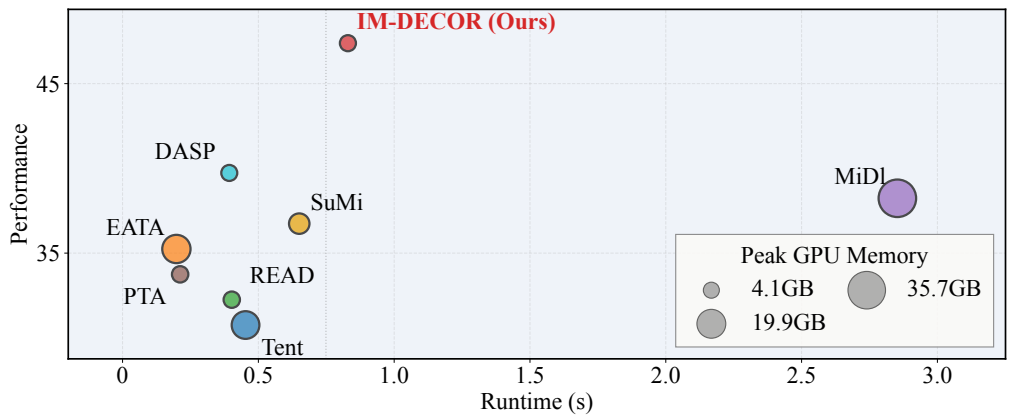


Figure 6: Efficiency comparison among IM-DECOR and baselines on the MM-IMDb dataset.

C Complexity and Efficiency Analysis

C.1 Complexity Analysis of IM-DECOR

In this section, we analyze the complexity of IM-DECOR, focusing on the G-DECOR and L-DECOR.

Definition 1. *To simplify the discussion, we consider a minimal TTA-MML setting in which each modality-complete instance consists of two modalities:* $X_i = \{x_i^{(1)}, x_i^{(2)}\}$*, and we take the complexity analysis on the decoration for modality* $m^{(1)}$ *as an example. Moreover, since MML-TTA is an online setting, we analyze the complexity on a given batch* $\mathcal{B}$*, without specifying the step. Let* N *denote the number of samples in the batch, and* $N^{(1)}$ *denote the number of samples that contain modality* $m^{(1)}$*. Let* K *represent the number of cluster prototypes maintained in the decoration container for modality* $m^{(1)}$ *(*$K \ll N^{(1)}$*). Let* d *denote the feature dimension of modality* $m^{(1)}$*.*

Complexity of G-DECOR. The G-DECOR consists of two stages: ❶ **Decoration Container Initialization**, and ❷ **Online Decoration Container Update**. For the modality $m^{(1)}$, the initialization applies K-means clustering on the representations $\mathcal{Z}^{(1)}$ with $N^{(1)}$ samples. Assuming I clustering iterations, the complexity is $\mathcal{O}(I \cdot N^{(1)} \cdot K \cdot d)$. For container update, given each incoming sample with modality $m^{(1)}$, we compute its distance to K prototypes and update the nearest one. The per-sample cost is $\mathcal{O}(K \cdot d)$, and therefore the batch-wise complexity is $\mathcal{O}(N^{(1)} \cdot K \cdot d)$.

Complexity of L-DECOR. The L-DECOR first performs instance-aware decoration via query generation and prototype retrieval. For each modality-missing sample, the projection network ϕ (MLP) maps the available modality representation to the query vector, with complexity $\mathcal{O}(d^2)$. The L-DECOR then computes similarity between the query and all K prototypes and performs weighted aggregation, resulting in $\mathcal{O}(K \cdot d)$. Moreover, the addition of learnable biases incurs a negligible cost of $\mathcal{O}(d)$. Therefore, the total per-sample complexity of L-DECOR is $\mathcal{O}(d^2 + K \cdot d)$.

C.2 Efficiency Analysis of IM-DECOR

In this section, we empirically evaluate the efficiency of IM-DECOR in terms of the run-time cost (s) and the peak value of GPU memory usage (GB), and compare it with baselines.

Specifically, we conduct the evaluation on the MM-IMDb dataset under the text-missing condition with a missing rate of 70%. As illustrated in Figure 6, we report the per-online-batch runtime cost (x-axis) and peak GPU memory usage (the size of the markers) with respect to performance, with batch size set to 128. We observe that IM-DECOR achieves modest runtime cost compared to the baselines. The peak GPU memory usage of IM-DECOR is *comparable to or even lower than several baselines* (e.g., Tent, EATA, MiDl), as most gradients are **only propagated** through the lightweight decoration modules, while bypassing the modality-specific encoder (often heavy for modern multimodal systems), avoiding the storage of large intermediate activations.

D MATHEMATICAL ANALYSIS OF IM-DECOR

In this section, we provide a theoretical analysis of our IM-DECOR from the perspective of Wasserstein-based domain adaptation (Shen et al., 2018; Ben-David et al., 2010), connecting the target-domain risk to a representation-level discrepancy.

Lemma 1 (Wasserstein Domain Adaptation Bound (Shen et al., 2018)). *Let each hypothesis $h \in \mathcal{H}$ be $\mathcal{K}$-Lipschitz continuous. Then the expected target-domain risk $\epsilon_T(h)$ is upper-bounded by*

$$\epsilon_T(h) \leq \epsilon_S(h) + 2\mathcal{K}\, W_1(\mu_S, \mu_T) + \lambda, \tag{22}$$

where $\epsilon_S(h)$ and $\epsilon_T(h)$ are the expected risks on the source and target domains, μ_S and μ_T are the feature-space distributions, $W_1(\mu_S, \mu_T)$ is the 1-Wasserstein distance between them, and λ is the combined error of the ideal hypothesis h^* that minimizes the combined error $\epsilon_S(h^*) + \epsilon_T(h^*)$.

Remark. The proof of this Lemma is in (Shen et al., 2018). For a fixed hypothesis space $\mathcal{H}$ and a fixed feature representation, λ is typically treated as a constant. Therefore, reducing the target-domain risk amounts to reducing the source risk together with the feature-level discrepancy term $W_1(\mu_S, \mu_T)$.

Lemma 2 (Natural Coupling under MML-TTA). *For clarity, we omit the time step and analyze MML-TTA at a given step after the adaptation parameters have been updated for the current batch and are treated as fixed. Consider a dual-modality setting, where $X = (x^{(1)}, x^{(2)}) \sim \pi$ is a random complete-modality sample, and $\widetilde{X} = (\varnothing, x^{(2)})$ denotes its modality-missing counterpart. Let $\mathcal{A}(\cdot)$ be a general adaptation method for adapted multimodal representation. Define the source- and target-domain representations under MML-TTA as:*

$$\mathbf{u}_S = H_\theta\big(f_\theta^{(1)}(x^{(1)}), f_\theta^{(2)}(x^{(2)})\big), \quad \mathbf{u}_T = \mathcal{A}\Big(H_\theta\big(\varnothing, f_\theta^{(2)}(x^{(2)})\big)\Big), \tag{23}$$

where $f_\theta^{(1)}$ and $f_\theta^{(2)}$ are the modality-specific encoders, H_θ is the multimodal fusion network. Let μ_S and μ_T denote the distributions of $\mathbf{u}_S$ and $\mathbf{u}_T$, induced by $X \sim \pi$. Then, $W_1(\mu_S, \mu_T)$ satisfies:

$$W_1(\mu_S, \mu_T) \leq \mathbb{E}_{X\sim\pi}\left[\|\mathbf{u}_S - \mathbf{u}_T\|_2\right]. \tag{24}$$

Proof. By definition, the 1-Wasserstein distance (Shen et al., 2018) is expressed as:

$$W_1(\mu_S, \mu_T) = \inf_{\gamma\in\Gamma(\mu_S,\mu_T)} \mathbb{E}_{(\mathbf{u}_S,\mathbf{u}_T)\sim\gamma}\left[\|\mathbf{u}_S - \mathbf{u}_T\|_2\right]. \tag{25}$$

Assumption: In the MML-TTA setting, the domain shift we focus on between the source and target mainly arises from modality missing rather than semantic changes (cf. Table 10 for evidence).

Consequently, $\mathbf{u}_S$ is constructed from the complete-modality input X, while $\mathbf{u}_T$ can be seen as being constructed from its modality-missing counterpart $\widetilde{X}$ via the adaptation. Since both $\mathbf{u}_S$ and $\mathbf{u}_T$ can be regarded as functions of the *same underlying random variable* $X \sim \pi$, they can induce a joint distribution γ_0 over $(\mathbf{u}_S, \mathbf{u}_T)$. By construction, the marginals of γ_0 are μ_S and μ_T, hence $\gamma_0 \in \Gamma(\mu_S, \mu_T)$. This coupling captures the intrinsic correspondence between complete and missing views of the same sample, which is a property unique to MML-TTA. By the definition of infimum:

$$W_1(\mu_S, \mu_T) = \inf_{\gamma\in\Gamma(\mu_S,\mu_T)} \mathbb{E}_{(\mathbf{u}_S,\mathbf{u}_T)\sim\gamma}\left[\|\mathbf{u}_S - \mathbf{u}_T\|_2\right] \leq \mathbb{E}_{(\mathbf{u}_S,\mathbf{u}_T)\sim\gamma_0}\left[\|\mathbf{u}_S - \mathbf{u}_T\|_2\right] \tag{26}$$

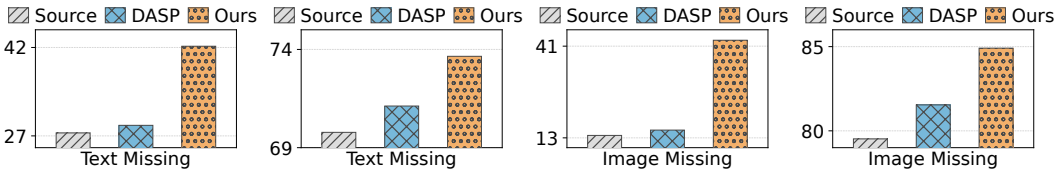
$$= \mathbb{E}_{X\sim\pi}\left[\|g_1(X) - g_2(X)\|_2\right] = \mathbb{E}_{X\sim\pi}\left[\|\mathbf{u}_S - \mathbf{u}_T\|_2\right], \tag{27}$$

which completes the proof of Lemma 2. □

Proposition 1 (Wasserstein Bound Induced by IM-DECOR). *Under the same setting, let $\hat{\mathbf{l}}^{(1)}$ denote the decorated representation constructed by IM-DECOR. Then the target-domain risk satisfies*

$$\epsilon_T(h) \leq \epsilon_S(h) + 2\mathcal{K}\,\mathbb{E}_{X\sim\pi}\left[\left\|H_\theta\big(f_\theta^{(1)}(x^{(1)}), f_\theta^{(2)}(x^{(2)})\big) - H_\theta\big(\hat{\mathbf{l}}^{(1)}, f_\theta^{(2)}(x^{(2)})\big)\right\|_2\right] + \lambda. \tag{28}$$

Proof. Under IM-DECOR, the $\mathcal{A}(\cdot)$ is instantiated by the G-DECOR and L-DECOR. Specifically, for a sample X in a given step after the adaptation parameters have been updated, the $\hat{\mathbf{l}}^{(1)}$ is determined by the remaining modality of this sample, and the resulting $\mathbf{u}_T$ can be treated as a function of X for the coupling analysis. Substituting this instantiation into Lemma 2 yields the result. □



(a) MM-IMDb. (b) Food101. (c) MM-IMDb. (d) Food101.

Figure 7: Performance of IM-DECOR and baselines under boundary MML-TTA scenarios.

Table 4: Statistics of benchmarks used in experiments. (S): Source Domain, (T): Target Domain.

Benchmark	Modality	#Class	#Train (S)	#Test (T)	Downstream Task
MM-IMDb	Text + Image	23	15,552	7,799	(Multi-Label) Movie Classification
Food101	Text + Image	101	61,142	22,716	Food Classification
Kinetics50	Video + Audio	50	29,204	2,466	Action Classification
VGGSound	Video + Audio	309	184,171	14,046	Sound Event Classification
IEMOCAP	Text + Audio + Video	4	2,717	938	Sentiment Classification
MOSI	Text + Audio + Video	7	1,284	686	Sentiment Classification

Remark. The IM-DECOR can be interpreted as attempting to reduce the $\epsilon_T(h)$ by decorating incomplete inputs to mimic their complete-modality counterparts in the representation space. Moreover, we empirically observe that IM-DECOR leads to improved alignment between the source and target representation distributions under MML-TTA (evidenced by Figure 5 and App. G.4).

E Boundary Behaviour of IM-DECOR

In this section, we analyze the behavior of our IM-DECOR under boundary MML-TTA scenarios. Specifically, to make the discussion concrete, below we consider two representative cases: ❶ a modality is entirely missing across several consecutive online batches ($\eta = 100\%$); ❷ the cold-start scenario, where a modality does not appear in the test-time data stream until a certain time point.

For ❶, the G-DECOR in IM-DECOR serves as a modality-level reservoir, which can still provide robust global, modality-wise decoration even when the incoming batches do not contain the specific modality, and the behavior of IM-DECOR remains stable (Figure 4 provides empirical evidence).

For ❷, when a modality does not appear until a certain time point (cold-start), the behavior of IM-DECOR degenerates to that of standard TTA baselines (e.g., leveraging entropy minimization to implicitly align the domain gap), as the environment effectively reduces to a modality-reduced setting. In this case, performance is inherently upper-bounded, since the gap caused by the missing modality cannot be effectively bridged without any observations from that modality. Nevertheless, once the information of this modality becomes available, IM-DECOR can launch effective adaptation, improving the robustness of the downstream models to the missing of this modality. Moreover, following prior works that relax the strict assumptions of TTA by allowing access to *minimal source-domain statistics* (rather than raw data) (Wang et al., 2024; Li & Feng, 2026), we provide an **alternative solution** for this scenario, where lightweight modality-wise prototypes from the source domain, if available, can be leveraged.

To validate this alternative solution for ❷, we further extend the cold-start scenario to its most extreme form on the MM-IMDb and Food101 datasets under text and image missing scenarios, where the missing rate η of a given modality reaches and remains at **the extreme level of** $\eta = 100\%$ throughout the entire test stream. We compare IM-DECOR with baseline DASP, a conventional TTA baseline that does not require complete modalities (baseline MiDl cannot handle this setting). Notably, we only assume that the lightweight modality-wise prototypes from the source domain are available, while the original training samples remain inaccessible, following (Wang et al., 2024). As illustrated in Figure 7, we observe that both the source model and DASP exhibit very poor performance, as this scenario inherently constrains the achievable performance to that of a modality-reduced setting. Nevertheless, when equipped with only a minimal amount of source-domain information, our IM-DECOR can stably adapt the downstream model to this boundary scenario, achieving superior performance.

F Detailed Experimental Setup

F.1 Benchmarks

To showcase the versatility of IM-DECOR in adapting a wide range of pre-trained downstream models to the test-time modality-missing instances, we conduct comprehensive evaluations across six benchmarks (cf. Table 4 for statistics): ❶ **MM-IMDb** (Arevalo et al., 2017), ❷ **Food101** (Wang et al., 2015), ❸ **Kinetics50** (Carreira & Zisserman, 2017), ❹ **VGGSound** (Chen et al., 2020), ❺ **IEMOCAP** (Busso et al., 2008), and ❻ **MOSI** (Zadeh et al., 2016), covering both dual- and tri-modality scenarios. We provide the detailed descriptions of these benchmarks below.

❶ **MM-IMDb**: MM-IMDb is a multimodal movie genre benchmark, where each sample consists of a movie plot summary and a poster image, and the task is to predict one or multiple genres for each movie from 23 labels (multi-label classification). We adopt F1-Macro as the metric.

❷ **Food101**: Food101 is a large-scale, 101-class multimodal food classification benchmark built on paired food images and recipe text. Compiled using Google Image Search, it inherently incorporates real-world noise and variability, presenting both challenges and opportunities for robust model development in food recognition tasks. We adopt Accuracy as the metric.

❸ **Kinetics50**: Kinetics50 is a 50-class action recognition benchmark derived from the Kinetics dataset (Yang et al., 2024), which is a comprehensive collection of YouTube videos, covering a diverse set of 400 distinct human action classes. This benchmark evaluates whether a method can exploit both motion patterns and acoustic context, which are often jointly informative for recognizing actions in unconstrained environments. We adopt Accuracy as the metric.

❹ **VGGSound**: VGGSound is a large-scale audio-visual benchmark. It consists of unconstrained video clips with aligned visual content and sounding events, and is commonly adopted for sound event recognition from video and audio modalities. Since many categories are characterized by subtle audio-visual correspondence in real-world scenes, this benchmark is well suited for evaluating robustness under modality missing in audio-visual recognition. We adopt the Accuracy metric.

❺ **IEMOCAP**: IEMOCAP is a multimodal conversational corpus. It provides transcribed speech, acoustic signals, and facial video from scripted and spontaneous dyadic interactions performed by actors. Following previous works (Tsai et al., 2019), four emotions (happiness, anger, sadness and neutral state) are selected for emotion recognition, and we leverage the Accuracy and F1 as metrics.

❻ **MOSI**: MOSI is a collection of opinion video clips. Each video is annotated with sentiment in the range [-3,3]. The dataset is annotated with labels for subjectivity, sentiment intensity, per-frame and per-opinion annotated visual features, and per-millisecond annotated audio features. We adopt the standard 7-class sentiment prediction setting (Guo et al., 2024), with Accuracy and F1 as metrics.

F.2 Source Model

In general, we do not impose strict assumptions on the architectures of downstream models for each task. Therefore, we assume that the source models follow a standard multimodal architecture (Radford et al., 2021; Kim et al., 2021), consisting of modality-specific encoders, a modality fusion network, and a classifier. Based on this, we construct the source models for each task, and the details of the modality data preprocessing and the source model structures of each multimodal task are below.

For the **MM-IMDb** and **Food101** benchmarks, following (Lee et al., 2023), we tokenize the text inputs with a maximum context length of 77 tokens, and decompose images into patches via resizing. We then adopt the pre-trained CLIP (Radford et al., 2021) text encoder and vision encoder as the modality-specific encoders, with 512-dimensional modality embeddings *without sequence dimension* as the encoded modality representations. We then use the concatenation of the text and vision embeddings as the fusion network, and use a two-layer MLP as the classifier.

For benchmarks **Kinetics50** and **VGGSound**, we follow (Yang et al., 2024) and extract audio-visual inputs from raw waveforms and video frames, where each audio clip is converted into a 1024×128 log-Mel filter bank and each video is represented by a single resized and normalized RGB frame. We adopt the CAV-MAE (Gong et al., 2022b) audio branch and vision branch as the modality-specific encoders, deriving 768-dimensional encoded modality-specific representations that *have the*

Table 5: Performance of different source models on the corresponding benchmark.

MM-IMDb	Food101	Kinetics50	VGGSound	MOSI		IEMOCAP	
F1-M	ACC	ACC	ACC	ACC	F1	ACC	F1
58.38	93.98	82.24	63.27	38.37	26.91	66.67	66.55

sequential dimension. We subsequently concatenate the two modality token sequences and feed them into the unified transformer blocks as the fusion network, and adopt an MLP as the classifier.

For benchmarks **IEMOCAP** and **MOSI**, we follow (Xu et al., 2024) and load the pre-extracted text, audio, and vision feature sequences from the sentiment pickle files provided by the prior studies (Guo et al., 2024). We implement three modality-specific MLP encoders to project the 300-dimensional text, 74-dimensional audio, and 35-dimensional vision features into 128-dimensional modality-specific representations, which *have the sequential dimension*. We then prepend a learnable summary token to each modality sequence, add modality embeddings and sinusoidal positional embeddings, and feed the concatenated multimodal token sequence into a 2-layer Transformer encoder as the fusion network. The three summary tokens are concatenated and further fed into a residual MLP classifier.

For each source model, we train it on the corresponding modality-complete training set, and select the best-performing model as the final source model (without any missing training strategies), which will be evaluated on the corresponding test set under the MML-TTA setting using various adaptation methods. As illustrated in Table 5, we report the performance of the source models on the corresponding test sets without modality missing. The results demonstrate that the models are well-trained (perform well on the variant of the target domain — modality-complete version of the test set), and that the distribution shifts (performance drops) primarily stem from modality missing.

F.3 Baselines

Since MML-TTA is a novel setting that requires adapting a source model, pre-trained on modality-complete data, to test-time modality-missing instances without redesigning or retraining the model, conventional MML methods are not suited for this setting. Consequently, we primarily compare IM-DECOR with competitive TTA baselines, including unimodal methods: Tent (Wang et al., 2021) and EATA (Niu et al., 2022), and multimodal TTA methods: READ (Yang et al., 2024), SuMi (Guo & Jin, 2025), MiDl (Ramazanova et al., 2025), PTA (Wang et al., 2025), and DASP (He et al., 2026b). Below, we provide the detailed descriptions and implementations of each baseline under MML-TTA.

Tent: Tent (Wang et al., 2021) is a fully TTA framework that improves robustness under distribution shifts by minimizing the entropy on unlabeled target batches and updating normalization statistics with channel-wise affine parameters online. We implement Tent under MML-TTA by updating the normalization layers in the pre-trained downstream models via the entropy minimization.

EATA: EATA (Niu et al., 2022) is an efficient anti-forgetting TTA framework that performs active sample selection for entropy minimization and constrains important parameters to alleviate catastrophic forgetting during online adaptation. Under MML-TTA, we implement it as an extension of Tent, and further filter the unreliable samples to skip the adaptation through the entropy threshold: $E_0 = 0.4 \ln C$, where C is the total number of classes in the given task.

READ: READ (Yang et al., 2024) is a multimodal test-time adaptation method designed to address multimodal reliability bias, where intra-modal distribution shifts induce information discrepancies across modalities, by learning self-adaptive inter-modal attention and a reliability-aware robust adaptation objective. Under MML-TTA, we follow its adaptation strategy by updating the first layer of the cross-modal attention module using its proposed confidence-aware loss function.

SuMi: SuMi (Guo & Jin, 2025) is a multimodal wild-TTA framework that stabilizes adaptation under mixed corruptions and missing modalities through interquartile-range smoothing, unimodal assistance for low-entropy sample identification, and mutual-information sharing across modalities. Under MML-TTA, we follow its setting by updating the normalization layers on the samples identified by its proposed interquartile range smoothing and unimodal assistance, using the mutual information sharing loss and the weighted entropy minimization objective.

MiDl: MiDl (Ramazanova et al., 2025) explores alleviating the modality-missing problem at test time in the task of egocentric video understanding, which encourages the model prediction to be invariant to the available modality by minimizing the mutual information, while using self-distillation to preserve the original complete-modality behavior. Under MML-TTA, we follow its setting by updating the normalization layers only on the modality-complete samples using its proposed objective, the combination of mutual information on each missing-pattern prediction and self-distillation.

PTA: PTA (Wang et al., 2025) is a multimodal TTA method that combats prediction bias under simultaneous multimodal domain shifts by partitioning target samples into reliable and unreliable subsets, reweighting them according to bias and confidence, and aligning attention patterns. Under MML-TTA, we follow its setting by updating the first layer of the cross-modal attention module with the combination of weighted entropy minimization and attention-guided alignment.

DASP: DASP (He et al., 2026b) is a multimodal TTA framework that decouples stability and plasticity by identifying the biased modality through feature redundancy and assigning asymmetric stable and plastic adapters to different modalities. Under MML-TTA, we follow its setting by adopting the plastic adapters on the missing modalities while applying the stable adapters on the remaining modalities, using the entropy minimization, diversity, and KL regularization.

For baselines, we strictly follow the optimal hyper-parameters (or their recommended ranges) reported in the original papers, and select the best-performing results under the MML-TTA setting. Notably, for some source models that do not include a cross-modal attention module, we update the last self-attention layer in the modality-encoder for baseline READ and PTA.

F.4 Implementation Details

We set the decoration container capacity $K = 4$ across all MML-TTA settings. Under extreme cases where the number of samples containing a given modality is smaller than K (e.g., extremely small online batch size), we temporarily set K to this number, retain these samples, and reinitialize the container once more samples become available in the coming test stream. During the adaptation, we use AdamW optimizer with a learning rate of 1×10^{-4}, weight decay of 10^{-2}. For MM-IMDb, Food101, IEMOCAP, and MOSI, the online batch size is set to 128, while for Kinetics50 and VGGSound it is reduced to 32 to fit within the memory constraints of the GPU, since the source models for them incurs more overhead. Following Lee et al. (2023); Ma et al. (2021), for baselines that do not modify the latent multimodal inputs, we impute missing modalities with zeros. The source model remains frozen during adaptation, and only the decoration parameters are optimized. The missing table (specifying which samples are missing) is sampled *independently* for each missing rate or type, and IM-DECOR and baselines are run three times under different missing tables, and the averaged results are reported as the final performance. All experiments use NVIDIA RTX 4090 GPU.

G Additional Experiments

G.1 Main Performance under MML-TTA across Other Missing Rates

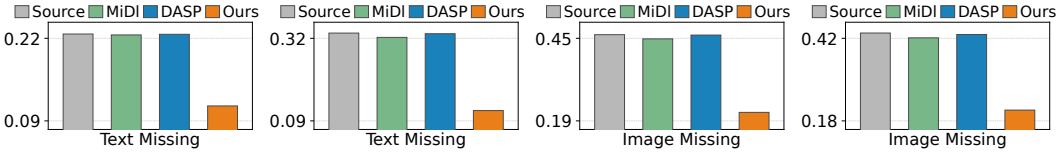
In the main paper, we conduct evaluation under MML-TTA by specifying the test-time modality missing rate η as 70%. In this section, we further provide the empirical results of IM-DECOR and baselines under 50% (medium) and 90% (extreme) missing rates, across all six benchmarks covering dual- and tri-modality scenarios. As presented in Table 6 and Table 7, we can draw a similar conclusion to that in main paper: IM-DECOR significantly outperforms all baselines across these additional missing rates, indicating its effectiveness under the MML-TTA setting.

G.2 Ablation Studies on Tri-modal Benchmarks

In the main paper, we conduct ablations on the dual-modality benchmarks. In this section, we further provide the ablative analysis on the tri-modal benchmarks, with results summarized in Table 8. We draw similar conclusions as in the main paper: each core component plays an important role in IM-DECOR, working hand-in-hand to achieve robust test-time modality-missing adaptation.

Table 6: Performance (%) of baselines and IM-DECOR on dual-modality benchmarks under 50% and 90% missing rates across diverse types. The best results are in **bold** and the second are underlined.

Methods	Venue	MM-IMDb				Food101				Kinetics50				VGGSound			
		T	I	Com.	Avg.	T	I	Com.	Avg.	A	V	Com.	Avg.	A	V	Com.	Avg.
Missing Rate = 50%																	
Source	N/A	46.15	41.65	44.05	43.95	81.99	86.63	84.30	84.31	73.24	62.81	67.84	67.96	42.65	56.33	50.44	49.81
Tent	ICLR'21	30.03	29.65	29.08	29.59	82.21	87.04	85.61	84.95	73.20	60.83	67.15	67.06	31.81	55.01	44.30	43.71
EATA	ICML'22	45.41	40.95	43.25	43.20	82.28	87.35	84.63	84.75	73.48	62.04	67.64	67.72	41.29	55.73	48.96	48.66
READ	ICLR'24	24.62	24.41	24.47	24.50	65.76	84.32	70.42	73.50	73.88	62.83	67.96	68.22	39.51	52.64	44.91	45.69
SuMi	ICLR'25	45.92	41.04	43.39	43.45	81.65	84.68	83.16	83.16	73.64	62.77	68.37	68.26	28.83	43.77	36.26	36.29
MiDl	ICLR'25	45.27	41.16	43.15	43.19	82.29	87.21	84.69	84.73	73.72	63.02	68.29	68.34	42.38	57.19	50.12	49.90
PTA	NeurIPS'25	28.16	28.48	27.33	27.99	66.05	71.90	72.83	70.26	54.87	58.64	56.00	56.50	35.64	48.18	42.73	42.18
DASP	CVPR'26	47.53	43.23	45.40	45.39	82.00	87.16	84.46	84.54	66.34	58.84	61.60	62.26	35.75	51.45	43.99	43.73
IM-DECOR	N/A	**51.80**	**51.91**	**52.04**	**51.92**	**83.80**	**89.36**	**86.68**	**86.61**	**75.34**	**64.19**	**69.10**	**69.54**	**44.95**	**57.73**	**51.75**	**51.48**
Missing Rate = 90%																	
Source	N/A	31.67	20.73	27.61	26.67	72.10	80.91	76.15	76.39	66.06	48.38	56.16	56.87	25.64	51.45	39.23	38.77
Tent	ICLR'21	18.69	13.03	16.87	16.20	72.62	80.99	77.58	77.06	65.82	45.26	55.84	55.64	08.26	50.09	29.87	29.41
EATA	ICML'22	31.80	20.90	27.64	26.78	72.65	82.29	76.85	77.26	65.82	48.66	56.81	57.10	26.23	51.14	39.10	38.82
READ	ICLR'24	25.43	25.36	21.21	24.00	58.94	84.07	65.78	69.60	66.87	48.95	56.81	57.54	22.75	46.19	33.77	34.24
SuMi	ICLR'25	31.17	19.58	26.94	25.90	71.82	79.41	76.25	75.83	66.42	48.50	56.45	57.12	13.00	43.32	28.69	28.34
MiDl	ICLR'25	31.61	21.71	27.67	27.00	72.39	81.94	76.81	77.05	66.50	48.82	56.73	57.35	24.78	52.15	38.62	38.52
PTA	NeurIPS'25	24.63	23.62	24.62	24.29	63.75	71.42	66.98	67.38	39.82	44.65	42.01	42.16	22.65	42.67	33.76	33.03
DASP	CVPR'26	34.20	24.11	29.03	29.11	72.36	81.97	76.37	76.90	57.42	47.12	50.08	51.54	24.61	49.37	33.13	35.70
IM-DECOR	N/A	**44.59**	**44.77**	**45.41**	**44.92**	**75.40**	**85.71**	**80.19**	**80.43**	**69.91**	**51.30**	**59.45**	**60.22**	**30.44**	**53.65**	**42.03**	**42.04**



(a) MM-IMDb. (b) Food101. (c) MM-IMDb. (d) Food101.

Figure 8: MMD performance of IM-DECOR and baselines under text or image missing scenario.

G.3 Additional Ablation Studies for IM-DECOR

To further disentangle the contribution of explicit decoration from generic entropy minimization (EM), we design two additional variants, with results reported in Table 9: ❶ **Decoration w/o EM**, where the decoration mechanism and online prototype update are retained while gradient-based optimization with $\mathcal{L}_{\text{EM}}$ is removed, and ❷ **EM w/o Decoration**, where the prototype-based decoration is removed while the learnable projection and residual parameters are optimized with $\mathcal{L}_{\text{EM}}$. We observe that *Decoration w/o EM* already substantially improves over Source, demonstrating that explicitly decorating the absent modality representation itself effectively alleviates the structural shift in MML-TTA. In contrast, *EM w/o Decoration* brings only marginal improvements or even degrades the Source performance, indicating that entropy minimization alone is insufficient to address such structural mismatch. Finally, the full IM-DECOR consistently outperforms both variants.

G.4 Quantitative Analysis of the Decoration using MMD

In the main paper, we have showcased the efficacy of IM-DECOR in bridging the domain gap under MML-TTA by decorating the modality-missing samples at test time through a visualization. We take a step further by introducing Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), a widely used metric for quantitatively measuring the representation-level distribution discrepancies between two domains. Mathematically, the definition of the standard MMD is given by (Gretton et al., 2012):

$$\text{MMD}^2(s,t) = \|\mathbb{E}_{x\sim s}[\phi(x)] - \mathbb{E}_{x\sim t}[\phi(x)]\|^2_{\mathcal{H}} \tag{29}$$

Table 7: Performance (%) of baselines and IM-DECOR on IEMOCAP and MOSI on 50% and 90% missing rates across diverse missing scenarios. The best are in **bold** and the second are underlined.

IEMOCAP		A		T		V		A-T		A-V		T-V		Avg.	
Methods	**Venue**	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Missing Rate = 50%															
Source	N/A	57.36	56.87	48.61	46.15	55.86	55.15	50.85	48.81	55.97	55.06	49.57	47.12	53.04	51.53
Tent	ICLR'21	55.97	54.67	45.95	43.16	55.12	54.21	49.47	46.03	54.90	53.07	47.01	43.96	51.40	49.18
EATA	ICML'22	55.86	54.90	47.44	44.72	55.33	54.45	49.89	47.08	54.58	53.14	47.76	44.92	51.81	49.87
READ	ICLR'24	50.96	50.74	46.06	45.51	50.75	50.95	46.27	45.80	50.75	51.00	44.78	44.13	48.26	48.02
SuMi	ICLR'25	56.40	55.47	48.40	45.71	55.97	55.43	49.25	46.05	55.33	53.95	50.21	47.29	52.59	50.65
MiDl	ICLR'25	55.86	55.09	47.23	44.54	55.44	54.91	49.79	47.52	54.26	53.03	48.29	45.88	51.81	50.16
PTA	NeurIPS'25	48.83	49.55	38.91	40.17	48.72	49.33	40.51	41.26	48.51	48.96	37.42	38.25	43.82	44.59
DASP	CVPR'26	38.17	35.24	25.80	24.15	36.89	33.26	26.65	24.30	38.06	34.80	26.65	24.83	32.04	29.43
IM-DECOR	N/A	**58.10**	**57.69**	**55.33**	**53.39**	**56.61**	**56.08**	**52.45**	**50.80**	**57.04**	**56.50**	**53.52**	**50.80**	**55.51**	**54.21**
Missing Rate = 90%															
Source	N/A	55.33	54.52	39.87	33.87	55.65	54.92	42.11	34.61	55.12	53.22	43.28	36.49	48.56	44.61
Tent	ICLR'21	55.97	54.12	36.57	30.01	54.26	53.26	41.90	33.18	52.45	48.83	40.30	33.85	46.91	42.21
EATA	ICML'22	55.82	55.46	38.59	32.62	54.37	53.58	40.94	32.79	53.52	50.76	41.58	35.27	47.47	43.41
READ	ICLR'24	50.53	50.52	44.67	39.68	50.11	50.28	37.53	34.86	51.92	52.12	46.70	39.93	46.91	44.57
SuMi	ICLR'25	55.33	53.83	42.54	35.82	54.37	53.93	43.07	34.84	54.69	52.52	46.59	38.79	49.43	44.96
MiDl	ICLR'25	56.08	54.85	40.72	34.39	53.84	53.26	43.60	35.93	54.05	51.63	43.28	36.37	48.60	44.41
PTA	NeurIPS'25	46.27	46.76	34.12	32.84	45.74	46.43	29.74	29.13	45.52	44.37	35.18	33.56	39.43	38.85
DASP	CVPR'26	37.63	34.71	16.74	10.28	36.99	33.02	16.74	09.65	36.57	32.22	16.63	09.00	26.88	21.48
IM-DECOR	N/A	**56.82**	**56.49**	**50.32**	**44.88**	**56.08**	**55.49**	**44.88**	**40.45**	**56.08**	**55.28**	**49.47**	**41.55**	**52.28**	**49.02**
MOSI		**A**		**T**		**V**		**A-T**		**A-V**		**T-V**		**Avg.**	
Methods	**Venue**	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Missing Rate = 50%															
Source	N/A	34.84	24.59	25.80	18.00	33.67	22.52	26.82	19.32	32.51	21.31	25.66	18.74	29.88	20.75
Tent	ICLR'21	31.20	21.58	23.03	15.54	30.17	19.95	25.36	17.49	27.84	17.71	26.82	18.02	27.40	18.38
EATA	ICML'22	31.05	21.70	23.62	16.37	32.36	21.59	25.66	17.88	29.88	19.34	26.38	18.06	28.16	19.16
READ	ICLR'24	33.97	22.92	24.34	16.07	34.99	22.59	25.95	18.33	32.51	20.78	26.53	17.28	29.72	19.66
SuMi	ICLR'25	34.26	24.56	24.78	17.60	33.53	23.42	26.53	19.15	31.20	20.56	25.51	17.81	29.30	20.52
MiDl	ICLR'25	32.80	23.16	24.78	17.91	33.24	23.04	25.95	18.55	29.45	19.30	26.68	17.97	28.82	19.99
PTA	NeurIPS'25	35.57	21.38	25.95	14.98	36.15	22.29	26.72	16.61	35.13	21.45	25.36	14.44	30.81	18.53
DASP	CVPR'26	27.11	11.08	21.87	09.95	28.28	11.72	23.03	09.95	27.41	11.05	24.64	10.72	25.39	10.75
IM-DECOR	N/A	**37.03**	**25.29**	**28.13**	**19.88**	**36.88**	**25.76**	**28.43**	**19.56**	**37.90**	**26.24**	**27.41**	**18.80**	**32.63**	**22.59**
Missing Rate = 90%															
Source	N/A	34.99	24.07	18.66	10.16	32.36	20.35	18.22	08.64	30.90	18.55	21.14	10.33	26.05	15.35
Tent	ICLR'21	31.05	20.56	17.64	08.47	28.72	17.98	17.20	07.27	26.09	15.18	21.14	04.99	23.64	12.41
EATA	ICML'22	31.34	21.67	17.93	09.05	31.34	20.07	17.78	07.95	29.01	17.79	20.11	11.49	24.59	14.67
READ	ICLR'24	33.67	23.09	18.80	11.00	35.28	22.38	18.37	09.14	32.22	21.52	19.39	11.73	26.29	16.48
SuMi	ICLR'25	33.67	23.58	18.08	09.66	30.76	20.11	17.35	07.64	28.86	17.37	18.66	10.69	24.56	14.84
MiDl	ICLR'25	32.94	22.96	17.78	09.20	31.49	20.30	17.93	08.60	27.84	16.93	20.12	10.24	24.68	14.71
PTA	NeurIPS'25	35.42	21.82	20.41	10.43	35.13	21.50	18.95	10.35	35.28	21.90	20.70	10.42	27.65	16.07
DASP	CVPR'26	27.11	10.30	19.39	08.42	27.99	10.66	18.22	07.88	26.97	10.01	20.26	11.54	23.32	09.80
IM-DECOR	N/A	**37.32**	**25.37**	**23.62**	**14.67**	**38.63**	**27.27**	**22.16**	**12.45**	**38.63**	**26.94**	**22.16**	**12.88**	**30.42**	**19.93**

Table 8: Ablation study (%) under 70% missing rate on tri-modality benchmarks.

		IEMOCAP						MOSI					
		A		T		V		A		T		V	
Module	**Variant**	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
G-DECOR	w/o Online	56.93	56.56	51.49	47.66	57.44	56.40	37.46	26.67	22.32	15.25	36.05	24.60
	w/o G-DECOR	56.46	56.05	47.91	44.59	57.10	56.30	36.03	26.33	21.45	14.31	34.03	22.55
L-DECOR	w/o Residual Bias	56.72	56.21	50.92	48.53	57.46	56.59	36.30	26.35	22.20	14.91	36.32	25.44
	w/o Regularization	57.25	56.84	51.56	48.36	57.68	56.86	36.76	25.89	23.76	15.68	36.76	25.34
	w/o L-DECOR	56.46	56.08	51.28	48.15	57.14	56.57	36.29	25.35	22.91	14.88	35.90	25.52
IM-DECOR	All	**57.68**	**57.27**	**52.56**	**49.36**	**58.10**	**57.44**	**37.76**	**26.89**	**24.05**	**16.11**	**37.17**	**26.33**

where $\phi(\cdot)$ maps samples into a RKHS space $\mathcal{H}$. Given two sets of samples $\mathcal{D}_s = \{x_i^s\}_{i=1}^M$ and $\mathcal{D}_t = \{x_j^t\}_{j=1}^N$ drawn from distributions s and t, the empirical estimate of MMD can be computed by expanding the original 2-norm formulation and applying the kernel trick (Yan et al., 2017):

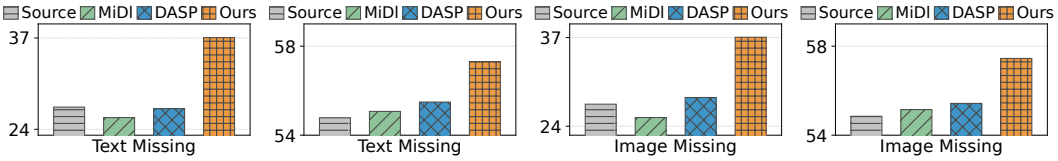
$$\text{MMD}^2(\mathcal{D}_s, \mathcal{D}_t) = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{i'=1}^{M} k(x_i^s, x_{i'}^s) + \frac{1}{N^2}\sum_{j=1}^{N}\sum_{j'=1}^{N} k(x_j^t, x_{j'}^t) - \frac{2}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} k(x_i^s, x_j^t), \tag{30}$$

Table 9: Additional ablation study of IM-DECOR.

Variant	MM-IMDb			Food101		
	T	I	Com.	T	I	Com.
Source	39.28	31.79	37.12	76.69	83.89	80.38
Decoration w/o EM	46.92	47.70	47.62	78.69	86.74	82.62
EM w/o Decoration	36.35	32.01	36.40	77.21	84.06	79.53
IM-DECOR	**48.22**	**48.88**	**48.99**	**79.32**	**87.76**	**83.76**

Table 10: MMD between samples from train and test set (under different missing types).

	Train-Test	Train-Test (Text Missing)	Train-Test (Image Missing)	Train-Test (Combined Missing)
MMD	0.00016	0.23537	0.46123	0.18851



(a) MM-IMDb. (b) Food101. (c) MM-IMDb. (d) Food101.

Figure 9: Performance of IM-DECOR and baselines on early-fusion ViLT backbone.

where $k(\cdot, \cdot)$ is a kernel function. In this work, we adopt the Gaussian kernel (Scholkopf et al., 1997):

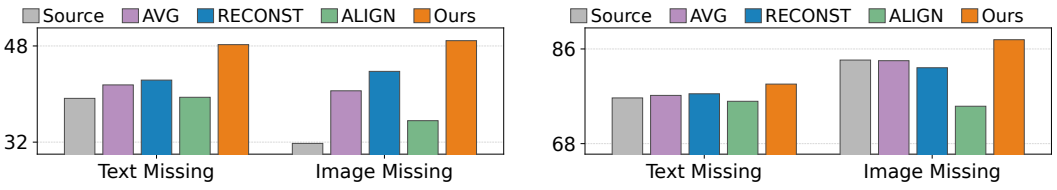
$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \tag{31}$$

where σ is the kernel bandwidth. Following common practice (Gretton et al., 2012), we set σ using the median heuristic, i.e., the median of pairwise distances among samples from the combination of two domains. We then measure the MMD under the MML-TTA setting, where the source domain is defined as the modality-complete version of the test set, and the target domain as its modality-missing counterpart. We conduct this evaluation on the MM-IMDb and Food101 datasets under text or image missing settings, where the target domain for each setting contains only image-only or text-only samples without any complete ones. The feature of each sample for MMD calculation is derived from the fused multimodal representation from the fusion network of the source model, after incorporating different adaptation methods. As illustrated in Figure 8, IM-DECOR effectively reduces the MMD compared to the original source model and all baselines, including MiDl and DASP.

In addition, after introducing the MMD metric, we measure the MMD between the training set and test sets under different missing types, aiming to analyze the main factors of source-target distribution shifts in the evaluation benchmark. Specifically, we conduct this analysis on MM-IMDb benchmark, leveraging the fused multimodal representations from the source model to calculate the MMD. The evaluation covers three missing types under a missing rate of 70%. As shown in Table 10, compared with the MMD between the original training and test sets, the MMD between the training set and the modality-missing test sets is several orders of magnitude larger. This indicates that, under our MML-TTA setting, the distribution shift between the source domain (training set) and target domain (test set with missing modalities) is dominated by modality missing.

G.5 Performance on the Early-Fusion Source Model

In main evaluation, we assess baselines and IM-DECOR on general and standard multimodal downstream models, where modality-specific encoders first transform raw modality data into latent representations, and a fusion network then performs cross-modal interactions over these representations. Such architectures are very common in real-world multimodal systems (Guo et al., 2025), where raw modality inputs are often unavailable or impractical to access directly, making encoder-based representation extraction followed by multimodal fusion a standard design choice. Nevertheless, in this section, we additionally evaluate competitive baselines and IM-DECOR on a representative



(a) MM-IMDb. (b) Food101.

Figure 10: Performance comparison among IM-DECOR and conventional MML baselines.

early-fusion model, ViLT (Kim et al., 2021), which has almost no modality-specific encoders and directly fuses raw modalities through a stack of attention layers. For IM-DECOR, since the decoration containers require modality-specific representations with meaningful semantic structure, we introduce a minimal modification on IM-DECOR to such structures. Specifically, we treat the first half of the fusion layers as modality-specific encoders, and use the output of the last layer in this half as the corresponding modality-specific representation. When deriving the modality-specific representations for a given modality, we also mask the remaining modalities before performing a forward pass to prevent interference. We conduct this evaluation on MM-IMDb and Food101 datasets under text and image missing types under 70% rate, and compare IM-DECOR with baselines MiDl and DASP. As illustrated in Figure 9, we observe that IM-DECOR still achieves the best performance on such structures, indicating its general MML-TTA capabilities across diverse downstream models.

G.6 Comparison with Conventional Modality Missing Methods

Although conventional MML methods are not directly applicable to the MML-TTA setting, as they rely on specific model architectures and well-designed training strategies with labeled data, we make appropriate modifications to representative conventional MML methods for MML-TTA and compare IM-DECOR with them. Specifically, we introduce three representative MML methods: ❶ **AVG** (Le Morvan et al., 2020): imputing the missing modalities with the averaged modality information from the data; ❷ **RECON** (Zhang et al., 2024): reconstructing the missing modalities using the remaining modalities and a well-designed reconstruction network; ❸ **ALIGN** (Li et al., 2024): aligning the representations of the fused modalities to the complete ones through a teacher-student distillation framework. For ❶ AVG, we simply leverage the averaged modality information from the test data stream to impute the missing modalities. For ❷ RECON, we introduce additional cross-modal reconstruction networks while avoiding internal interactions with the source model, in order to comply with the MML-TTA setting (i.e., without modifying the architecture of the source downstream model). The reconstruction network for each modality is implemented as a two-layer MLP and is optimized only on modality-complete samples observed at test time. For ❸ ALIGN, to comply with the MML-TTA setting, we do not adopt sophisticated teacher–student distillation frameworks that require modifying the source model. Instead, we align the fused representations of modality-missing samples with those of modality-complete ones, and perform adaptation by optimizing the normalization layers of the modality encoders and fusion network in the source model.

We compare the three methods with IM-DECOR on the MM-IMDb and Food101 benchmarks under text or image missing scenario with missing rate at 70%. As illustrated in Figure 10, comparing to our IM-DECOR, all three conventional MML baselines exhibit weak performance under the MML-TTA setting, and even underperform the source model in some modality-missing cases. For the ❶ baseline AVG, although it leverages modality-level information to mitigate the domain gap, such information is relatively coarse-grained compared to the prototype-based decoration containers in IM-DECOR, and thus lacks the ability to be continually refined over the test stream. Moreover, it ignores instance-level variations when performing the filling, failing to account for the diverse characteristics across samples. Finally, it does not introduce additional learnable components for adaptation, and thus cannot adapt the source model beyond the observed test data. The ❷ baseline RECON exhibits very limited performance, as the test data arrive in an online, batch-wise manner, and there are insufficient samples to effectively train the reconstruction networks. This issue is further exacerbated as the missing rate increases, where the model can only rely on a small number of modality-complete samples for optimization. Moreover, in the absence of label supervision, the cross-modal reconstruction signals become unreliable, leading to limited effectiveness. For the ❸ baseline ALIGN, its effectiveness is also limited under the MML-TTA setting due to the scarce

Table 11: Comparison (%) with unimodal models under a 70% missing rate.

	MM-IMDb		Food101	
Method	**T**	**I**	**T**	**I**
Unimodal	45.48	46.67	77.46	85.48
IM-DECOR	**48.22**	**48.88**	**79.32**	**87.76**

Table 12: Statistics of the similarity scores from L-DECOR. Top-1 M: Top-1 Mean.

	Text		Image		Both-Text		Both-Image	
	Top-1 M	$P(\text{top-1} > 0.5)$	Top-1 M	$P(\text{top-1} > 0.5)$	Top-1 M	$P(\text{top-1} > 0.5)$	Top-1 M	$P(\text{top-1} > 0.5)$
Value	0.3608	7.1%	0.4119	17.1%	0.3201	8.0%	0.3860	10.8%

modality-complete data under high missing rates (like baseline RECON), leading to insufficient supervision for reliable alignment. Moreover, since the source model is already well-trained, updating its parameters during adaptation may disrupt its performance on the corresponding task.

G.7 Comparison with Unimodal Method

We further compare IM-DECOR with a unimodal baseline. Specifically, we construct it by removing the other modality-specific branch from the source model and training it with supervised unimodal data. At test time, the resulting model is directly evaluated on the corresponding available modality, with results reported in Table 11. We observe that IM-DECOR consistently outperforms this unimodal alternative for three reasons. (1) Although a unimodal model avoids the distribution shift caused by modality missing, it is inherently restricted to the information carried by a single modality. In contrast, IM-DECOR decorates the missing representation and enables the pretrained multimodal model to recover and exploit complementary multimodal information even for incomplete samples. (2) Unimodal solutions scale poorly in practice, as a separate model must be trained and maintained for each possible modality subset, with the number of models rapidly increasing as more modalities are involved. (3) For modality-complete samples, unimodal models necessarily discard available complementary modalities, leading to substantial performance degradation, whereas IM-DECOR retains the original multimodal model and its capability on complete inputs.

G.8 Further Analysis on L-DECOR for Instance-Aware Decoration

In the previous empirical evaluations, we ablate the L-DECOR to showcase its important role. In this section, we further analyze the effectiveness of L-DECOR in providing instance-aware decorations, rather than degrading to simply averaging the prototypes in the decoration containers for diverse samples. Specifically, we report the statistics of the similarity scores between the query constructed from the remaining modality of an incomplete sample and the prototypes in the decoration containers, using the MM-IMDb benchmark under three modality-missing scenarios with a 70% missing rate. The suffixes "Both-Text" and "Both-Image" indicate the missing modality of the samples used for calculation. As presented in Table 12, compared with the uniform assignment of 0.25 similarity to each prototype, the mean similarity (average results on the whole target domain incomplete samples) between the query and the Top-1 activated prototype (Top-1 M) exceeds 0.32, indicating that L-DECOR does not degenerate into simple average pooling. Moreover, a non-negligible proportion of samples exhibit a Top-1 similarity score above 0.5, further confirming the efficacy of L-DECOR.

H Discussion of MML-TTA, Design of IM-DECOR, and Baselines

H.1 Why IM-DECOR is Model-Agnostic and Its Boundary

We use *model-agnostic* to denote independence from model-specific architectural redesign or retraining, rather than black-box access. IM-DECOR only requires a generic multimodal interface that exposes modality-specific latent representations and allows decorated representations to be injected before multimodal fusion. Such an encode-then-fuse interface is shared by a broad family of

multimodal models, allowing IM-DECOR to serve as a plug-and-play adaptation wrapper without modifying the original architecture or pretrained parameters. Its broad applicability is further validated across different modality compositions, tasks, and source-model architectures. IM-DECOR does require access to intermediate modality representations.

H.2 What Fundamentally Distinguishes IM-DECOR from Existing TTA Methods

The distribution shift in MML-TTA is fundamentally different from those considered in conventional TTA (including multimodal TTA) (Wang et al., 2021; Yang et al., 2024). Existing unimodal and multimodal TTA methods mainly address corrupted or unreliable inputs, where the corresponding modality and its representation still exist, i.e., $z^{(m)} \rightarrow \tilde{z}^{(m)}$. In contrast, modality missing causes the entire modality representation to disappear, i.e., $z^{(m)} \rightarrow \varnothing$. Therefore, adapting an existing shifted representation, as commonly done in TTA, is insufficient when there is no representation to adapt. IM-DECOR is designed **precisely for this structural shift**. Instead of modifying the downstream model to accommodate an incomplete input, G-DECOR collects modality-level information from the test stream, while L-DECOR converts such global resources into instance-specific decorations conditioned on the observed modalities. In this sense, conventional TTA primarily adapts the model to shifted inputs, whereas IM-DECOR adapts incomplete inputs to the pretrained model.

H.3 Why SuMi and MiDl are Limited under MML-TTA

For baseline **SuMi** (Guo & Jin, 2025), it treats modality missing as one type of strong OOD shift within the broader multimodal wild-TTA setting, and replaces a missing modality with zero-valued inputs. Its IQR smoothing controls when strongly shifted samples participate in adaptation, but does not resolve the absent modality representation itself. Its unimodal-assisted selection relies on predictions from individual modalities, whose intended interpretation becomes unreliable when one branch contains no valid modality observation. Similarly, its mutual-information sharing requires informative predictions from different modalities, while a missing branch provides no genuine modality information to share. Hence, SuMi mainly stabilizes adaptation around an incomplete input rather than explicitly repairing its structural absence. For baseline **MiDl** (Ramazanova et al., 2025), it addresses test-time missing modalities in downstream task of *egocentric video understanding*, but its adaptation relies on modality-complete samples. As the missing rate increases, fewer incoming samples can provide effective adaptation signals, and the method becomes particularly restricted when complete samples are absent. Moreover, MiDl updates the source model during adaptation, which may degrade its performance on in-domain complete samples, whereas IM-DECOR keeps the downstream model frozen and adapts only lightweight decoration parameters. Consequently, IM-DECOR can maintain the performance on complete samples.

H.4 Why MML-TTA Assumes Stream-Level Modality Coverage

MML-TTA allows any modality to be missing for individual samples, entire batches, or even consecutive segments of the test stream. We only assume that every modality is observed at least occasionally over the entire stream. This assumption **differs from conventional MML**, where models can learn cross-modal structure and missing-modality robustness during source training from modality-complete and synthetically masked data. In MML-TTA, however, neither the source data nor the training process is accessible, and adaptation must rely solely on the target stream. If one modality never appears throughout the entire stream, no target-side observation is available to characterize it, and a purely test-time method cannot recover information that is never observed without introducing additional source statistics, generative priors, or external knowledge. The problem then degenerates from intermittent modality missing into a permanently modality-reduced setting. Importantly, this assumption does not require any modality-complete sample. Once a modality has been observed, G-DECOR can retain its accumulated information and continue supporting decoration even when that modality subsequently disappears for long periods.

I Limitation & Future Work

Despite the strong effectiveness and robustness of IM-DECOR across diverse modality-missing scenarios, some important aspects remain underexplored. In particular, explicitly characterizing the

uncertainty of decorated missing-modality representations may help assess their reliability and enable more robust uncertainty-aware adaptation under severe missing conditions.

I.1 Extension of MML-TTA Research Problem

In this work, we focus on addressing the domain shift induced by modality missing between the source and test domains, which aligns with the standard setting in multimodal learning. While this focused formulation enables a clear investigation of the problem, it also leaves several promising directions for future research. Specifically, current framework IM-DECOR focuses on handling structural shifts induced by modality missing, which is a well-established and widely accepted setting in multimodal learning and test-time adaptation, and is also commonly encountered in real-world scenarios. Nevertheless, in more extreme cases, test-time data may exhibit both modality missing and additional semantic distribution shifts to a certain degree. Extending IM-DECOR to jointly handle such compounded shifts would require designing additional mechanisms and adaptation strategies.

I.2 Extensible Design of DECOR with Potential Future Direction

In IM-DECOR, the two decorators, G-DECOR and L-DECOR, can be highly abstracted into two functional components: ❶ a global decorator that aligns the latent representations of missing modalities with the corresponding modality spaces using modality-level global statistics, and ❷ a local decorator that further refines these representations at the instance level to achieve sample-specific adaptation. This abstraction highlights the intrinsic extensibility of the IM-DECOR: with more complex designs, these modules could be adapted to handle more specialized or extreme scenarios. Nevertheless, in this study, we adopt a **minimalist design** to accommodate a broad range of application scenarios, while leaving the following directions for future extensions in highly customized situations.

Potential Extension of G-DECOR. The global decorator could be extended to better handle challenging streaming scenarios. For instance, in highly imbalanced or temporally correlated test streams (Gong et al., 2022a), techniques commonly used in TTA, such as pseudo-labeling-based filtering (Gong et al., 2022a; Liang et al., 2020), could be incorporated to stabilize the global modality distributions and prevent prototype skewing during K-means update, allowing G-DECOR to maintain robust global representations under extreme conditions.

Potential Extension of L-DECOR. The local decorator could be enhanced to exploit the remaining modalities more effectively, but accompanied with increased overhead. One possible approach is to leverage large pre-trained models to generate richer representations of the available modalities, enabling potentially better adaptation in complex modality-missing scenarios.

J Broader Impacts

This work introduces a novel MML-TTA setting with an effective framework IM-DECOR, with a key advantage that it can be directly integrated with existing, already-deployed multimodal systems **without requiring any modification** to their architectures or training procedures. Such a plug-and-play property is enabled by the fact that our method operates at the input level, treating inputs as **a universal test-time modality-missing adaptation anchor** that is shared across different downstream multimodal models. As a result, IM-DECOR can be applied to a wide range of pre-trained systems regardless of their internal designs. This makes our method particularly suitable for real-world applications, where models are often pre-trained and deployed, and redesigning or retraining them is costly or impractical. By enabling seamless adaptation to missing modalities at test time, our method improves the robustness and reliability of multimodal systems under imperfect conditions, such as sensor failures, privacy constraints, or communication limitations. This capability is especially valuable in safety-critical domains, including healthcare and autonomous driving, where maintaining stable performance under incomplete inputs is essential. At the same time, improving robustness under modality-missing scenarios may contribute to more resilient and accessible AI systems.