TTSD-FAR: TEST-TIME SELF-DISTILLATION WITH FISHER-ANCHORED RESTORATION FOR ADAPTATION OF LVLMS UNDER MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

LVLMs have shown remarkable performance on multimodal tasks like multimodal emotion recognition (ER) in-the-wild. ER is inherently multimodal, requiring a joint understanding of facial expressions, vocalizations, language, biosignals, and gestures. However, real-world deployment remains challenging: modalities may be missing or noisy at test time. Partial observations can be viewed as a distribution shift relative to the complete-modality distribution. SOTA TTA methods based on entropy or perplexity do not transfer to autoregressive LVLMs, while RAG degrades when the observed modality is weak. Because no ground-truth supervision exists to verify individual updates, adaptation across this stream risks accumulating drift and degrading once the model departs from a reliable solution. An effective solution must therefore adapt to missing modalities and remain effective during continual adaptation. We address both jointly with **Test-Time Self-Distillation (TTSD)**, a parameter-efficient framework in which a frozen teacher, trained on complete modalities, guides an adaptive low-rank student via self-distillation, updating only a negligible number of parameters. Stability is built into this same loop through **Fisher-Anchored Restoration (FAR)**, which monitors Fisher information stability to detect convergence versus drift and restores the student to the ACTIVE state when distributional shifts are identified. Our experiments[1] on MELD, DFEW, and BAH under 0%–50% missing modalities show that this unified adaptation-restoration design consistently outperforms entropy-based adaptation, RAG, and perplexity-based generation.

1 INTRODUCTION

The rapid advancement of LVLMs Lin et al. (2023) has revolutionized multimodal learning tasks, achieving unprecedented performance. These models leverage transformer architectures with billions of parameters, pre-trained on massive video-text corpora, to develop rich cross-modal representations Radford et al. (2021). LVLMs are promising for video-based multimodal ER (MER), where systems typically capture facial, vocal, and language cues to address subtle (or compound) expressions and high inter-subject variability Huang et al. (2025). Recent work shows that LVLMs can capture fine-grained emotional and health states and contextual relationships that traditional discriminative models fail to detect Ge et al. (2024). However, their deployment in real-world applications confronts a fundamental challenge that has received surprisingly little attention: missing modality at test time.

In practical video-based MER scenarios, modality absence is ubiquitous and inevitable. Inputs may suffer from audio corruption in noisy environments, transcription services may fail due to poor audio quality or privacy constraints, and sensor malfunctions can eliminate entire modalities during data collection Wang et al. (2020). Traditional approaches address this through training-time strategies, either by training separate models for each modality combination or by employing modality dropout during training Nezakati et al. (2024). Yet both approaches are impractical for LVLMs, as retraining billion-parameter models for each missing modality scenario is computationally prohibitive (requiring 10^{20} FLOPs for 7B parameter models), while training-time modality dropout

[1] Our code is included in the supplementary materials and will be made public.

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.

degrades full-modality performance and cannot adapt to deployment-specific missing patterns (Ma et al., 2021). Ramazanova et al. (2025) pose the missing modality as a domain shift problem and propose a method for TTA. The authors claim that despite advancements in the missing modality literature, a common drawback persists: all state-of-the-art approaches necessitate expensive retraining of the multimodal model. This poses a substantial challenge, particularly in applications with: (*i*) extensive training data; (*ii*) where the retraining process is prohibitively expensive; and (*iii*) when source data is not available, making the aforementioned approaches impractical.

As shown with recent foundation models (e.g., GPT-style architectures and large multimodal systems Maaz et al. (2024)), retraining large language models (LLMs) and LVLMs is prohibitively expensive, requiring specialized hardware and weeks of distributed optimization. It is therefore impractical to retrain or fully fine-tune these models for every downstream task or deployment setting due to catastrophic forgetting, computational cost, and source data availability. This motivates the development of lightweight TTA methods for efficient customization without modifying the full parameter set.

We investigate the domain shift caused by missing modality inputs from 3 different viewpoints: (*i*) UMAP visualization of the model's internal representations; (*ii*) MMD computations between the full and missing modality inputs of the same sample; and (*iii*) the degradation of performance. We perform this analysis on the MELD dataset with 50% text-missing inputs using the *Video-LLaVA-7B* model. Figure 1 displays the representation geometry of the complete and text-missing inputs. For visualization, we select two emotion classes from MELD; the two colors correspond to these classes. We use this two-class subset solely for visualization clarity. The • represents the full modality inputs and $\times$ represents the missing modality input for the same data samples. It reveals that the full and missing inputs are not only offset, but that missing inputs also blur the decision boundary. With full-modality inputs, the two classes form well-separated clusters, whereas under partial observation, this separation collapses entirely, with both classes overlapping into a single indistinguishable region

Quantitative analysis reveals a substantial internal distribution shift when modalities are missing, with an MMD of 0.8350 and cosine similarity of 0.6342 between complete and missing inputs for the same input samples. This indicates that the absence of modalities can significantly disturb the learned attention manifold, even though the semantic content remains unchanged. Lastly, the decline in performance shown in Tables 1-3 indicates that LVLMs experience a severe performance drop (F1-score 0.6149 $\rightarrow$ 0.4785) in the 50% text-missing case. Entropy-based TTA methods tend to drop towards near-random performance after adaptation. This observation is in line with the findings of Hu et al. (2025), which showed that entropy minimization objectives are ill-suited for the TTA of LLMs.

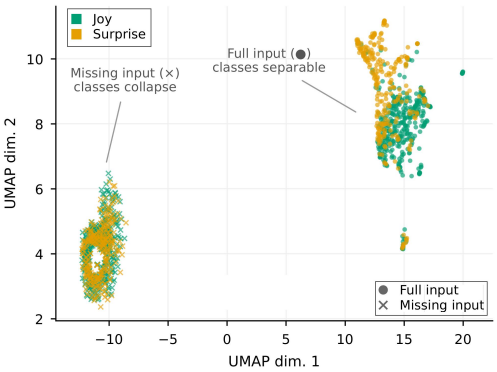

Figure 1: UMAP of the internal representations. The cross ($\times$) represents the missing modality inputs, and the circle (•) represents the full modality inputs.

A second, less-addressed challenge in online TTA for LLMs and LVLMs is perpetual adaptation. In practice, two failure modes emerge. First, once the student LoRA has converged, continued updates introduce gradient noise that perturbs the learned representation without improving performance Wang et al. (2022); Niu et al. (2023). Second, test streams in real deployments are non-stationary: speaker identities, recording conditions, and missing-modality rates shift over time. Perpetual adaptation without drift awareness risks overwriting consolidated knowledge, structurally analogous to catastrophic forgetting Kirkpatrick et al. (2017). Monitoring the raw distillation loss is an unreliable criterion for either stopping or restarting: the loss can plateau due to sample difficulty rather than genuine parameter convergence and can spike due to hard within-distribution samples rather than a true distributional shift. This motivates two design choices: adaptation needs a supervision signal that fits autoregressive LVLMs, motivating latent alignment over entropy. A stopping signal grounded in the parameters themselves rather than the loss, motivating the Fisher information diagonal, whose shift from its value at convergence exposes genuine drift and lets adaptation pause and reactivate accordingly.

The contributions of this paper are summarized as follows: **(1)** TTSD: a parameter-efficient TTA method tailored to autoregressive LVLMs under missing modalities, built around conditional latent representation alignment as a supervision signal that overcomes the failure modes of entropy-based objectives and retrieval methods. **(2)** FAR: a two-state stopping and drift-detection and reactivation mechanism for TTSD. It monitors the Fisher information diagonal to detect student convergence (ACTIVE $\rightarrow$ ANCHORED) and uses a Fisher Mismatch Index to detect test-stream distributional shift (ANCHORED $\rightarrow$ ACTIVE). **(3)** An extensive set of experiments on three challenging video-based ER datasets, namely MELD Poria et al. (2019), DFEW Jiang et al. (2020), and BAH González-González et al. (2026), with multiple missing ratios and across different modalities shows the effectiveness of our proposed TTSD-FAR method.

2 Related Works

Learning with Missing Modalities. Multimodal learning with missing modalities has been extensively studied, with early works such as ModDrop Neverova et al. (2016), SMIL Ma et al. (2021), and masked modality projection (MMP) Nezakati et al. (2024) introducing train-time strategies to enhance robustness. Maheshwari et al. (2024) propose a multi-modal teacher for masked modality learning, improving semantic segmentation under missing modalities. Some methods have explored modality-invariant architectures where single-branch models trained to be invariant to modality absence exhibit robustness during both training and testing Saeed et al. (2024). Guo et al. (2024) use prompt learning in specialized cross-modal transformer architectures to regenerate the missing modality. More recently, Chen et al. (2026) proposed SMCIR, which detects sample-level modality missingness via an unsupervised entropy/mutual-information/similarity score (DMFD), then reconstructs missing modalities through context-guided multi-scale cross-modal attention (CMCG). Despite this progress, most methods rely on training-time strategies, either reconstructing missing modalities, simulating modality absence during training, or making custom architectural changes. These train-time methods are not directly transferable to LVLMs, which are expensive to re-train and whose generative decoders lack clear self-supervised objectives that correlate with generation quality under modality absence.

Test-Time Adaptation. TTA and test-time training have emerged to handle distribution shifts at inference time without access to target labels Wang et al. (2022). TENT proposes entropy minimization by adapting normalization statistics at test time Wang et al. (2021). EATA Niu et al. (2022) improves entropy-based TTA by selectively updating model parameters using only reliable low-entropy samples and regularizing updates with a Fisher constraint to prevent forgetting. Many recent works analyze batch normalization for robustness under shift Nado et al. (2021). A related line of continual-TTA work decides when to reset via drift detection rather than fixed schedules or entropy alone: periodic resets to pretrained weights curb long-horizon collapse Press et al. (2023). Mishra Mishra (2026) introduced RDumb++, a principled extension of RDumb that introduced two drift-detection mechanisms, i.e., entropy-based drift scoring and KL-divergence drift scoring, together with adaptive reset strategies. Alternating domain construction was proposed by Lin et al. (2024). Sun et al. (2020) instead optimize auxiliary self-supervised losses during inference to improve adaptation. These methods are tailored to models that produce a single softmax distribution over a small, fixed label set, where output entropy is a direct, low-dimensional measure of prediction confidence. Self-supervised objectives and associated stopping signals do not transfer to LVLMs. Hu et al. (2025) introduced test-time learning for LLMs (TLM), adapting LLMs during inference using only unlabeled test data via input perplexity minimization rather than entropy minimization or supervised fine-tuning. Adapting entire layers or normalization statistics in billion-parameter models further raises stability and efficiency concerns, motivating lightweight, parameter-efficient adaptation instead. Unlike EATA's per-step Fisher regularization, TTSD-FAR uses the Fisher diagonal as a control signal that suspends and resumes adaptation entirely. Ramazanova et al. (2025) propose MiDl, a test-time adaptation method tailored to missing modalities that minimizes mutual information between the prediction and the available modality while distilling from the original model; however, it is designed for lightweight discriminative backbones and estimates this objective via separate forward passes over a small, enumerable set of modality combinations, an assumption that breaks down for billion-parameter autoregressive LVLM decoders.

Retrieval-Augmented Generation. RAG methods improve LLM and LVLM's performance during inference by incorporating external knowledge. Relevant information from a knowledge base is retrieved during inference Asai et al. (2023); Zhao et al. (2024); Qian et al. (2025). RAG meth-

ods prove effective for tasks that require domain knowledge, but they rely heavily on the quality of retrieved information and incur additional computational latency. Precomputed vector databases or memory banks are also required for retrieval during inference Abootorabi et al. (2025). State-of-the-art methods to address missing modalities either require retraining with each modality combination or depend on discriminative learning objectives and specialized architectures to hallucinate missing modalities. The main issue faced by LVLMs for TTA is that there is no strong supervision signal available for adaptation. Entropy objectives do not provide meaningful supervision, and RAG schemes fail if the observed modality is too weak to retrieve correct class neighbors. This paper fills such a gap by proposing a novel approach for feature-level adaptation using dual LoRA adapters through self-distillation and FAR for convergence monitoring that allows stopping and resuming adaptation dynamically.

3 Proposed Methodology

Let $\mathcal{X} = \mathcal{X}^{\text{v}} \times \mathcal{X}^{\text{t}}$ denote the multimodal input space consisting of video and text modalities, where $\mathcal{X}^{\text{v}}$ represents video frames and $\mathcal{X}^{\text{t}}$ represents textual utterances. Let $\mathcal{Y}$ denote the output label space (e.g., emotion classes). We assume access to a pre-trained LVLM $f^{\theta} : \mathcal{X} \to \mathcal{Y},\ \theta \in \mathbb{R}^d$, trained on large-scale multimodal corpora. The model consists of (i) modality-specific encoders $E^{\text{v}} : \mathcal{X}^{\text{v}} \to \mathbb{R}^{n \times d^{\text{v}}}$, $E^{\text{t}} : \mathcal{X}^{\text{t}} \to \mathbb{R}^{m \times d^{\text{t}}}$, and (ii) a LLM$^{\theta}$ backbone : $\mathbb{R}^{(n+m)\times d} \to \mathcal{Y}$, which performs cross-modal fusion and autoregressive reasoning. We denote the frozen pre-trained backbone parameters by θ^0 and the teacher and student LoRA weights by $\Delta\theta^{\text{tea}} = B^{\text{tea}}A^{\text{tea}}$ and $\Delta\theta^{\text{stu}} = B^{\text{stu}}A^{\text{stu}}$, respectively. We further denote the vectorized student LoRA parameters as $\phi = \text{vec}(A^{\text{stu}}, B^{\text{stu}}) \in \mathbb{R}^p$, where p is the total number of student LoRA parameters.

3.1 Problem Formulation

3.1.1 Latent Inference under Partial Observation.

Let $x = (x^{\text{t}}, x^{\text{v}}) \sim P_{\text{full}}$ denote complete multimodal inputs drawn from the training joint distribution. At test time, modalities may be missing. We introduce binary modality masks $m^{\text{v}}, m^{\text{t}} \in \{0, 1\}$, and define the partially observed input as

$$\tilde{x} = (\tilde{x}^{\text{v}}, \tilde{x}^{\text{t}}) = (m^{\text{v}} \cdot x^{\text{v}},\ m^{\text{t}} \cdot x^{\text{t}}). \tag{1}$$

The teacher's intermediate representation at layer l is

$$h^{\text{tea}}(x) = \text{LLM}_l^{\theta^0+\Delta\theta^{\text{tea}}}(x^{\text{v}}, x^{\text{t}}) \in \mathbb{R}^{s\times d}. \tag{2}$$

The missing-modality setting can be viewed as a *partial observation problem*, where the learner must infer a representation consistent with the latent multimodal structure learned from complete data. We interpret $h^{\text{tea}}(x)$ as a latent variable lying on a multimodal representation manifold $\mathcal{Z}$. Under squared error loss, the Bayes-optimal estimator of the complete-modality representation given the observed input $\tilde{x}$ is the conditional expectation

$$h^*(\tilde{x}) = \mathbb{E}\left[h^{\text{tea}}(x) \mid \tilde{x}\right]. \tag{3}$$

3.1.2 TTSD as Conditional Latent Reconstruction.

The student adapter parameterized by $\Delta\theta^{\text{stu}}$ produces

$$h^{\text{stu}}(\tilde{x}) = \text{LLM}_l^{\theta^0+\Delta\theta^{\text{stu}}}(\tilde{x}^{\text{v}}, \tilde{x}^{\text{t}}) \in \mathbb{R}^{s\times d}. \tag{4}$$

TTSD minimizes the feature-level distillation objective

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{x\sim P_{\text{full}}}\left[\left\|h^{\text{stu}}(\tilde{x}) - h^{\text{tea}}(x)\right\|_2^2\right], \tag{5}$$

whose minimizer satisfies $h^{\text{stu}*}(\tilde{x}) = \mathbb{E}[h^{\text{tea}}(x) \mid \tilde{x}]$, so that under a squared-error objective, the student is encouraged to approximate the conditional expectation of the teacher representation given partial input.

However, perpetual adaptation of ϕ across an unbounded test stream introduces two failure modes: (i) once $\Delta\theta^{\text{stu}}$ has converged, continued gradient updates add noise without improving the learned

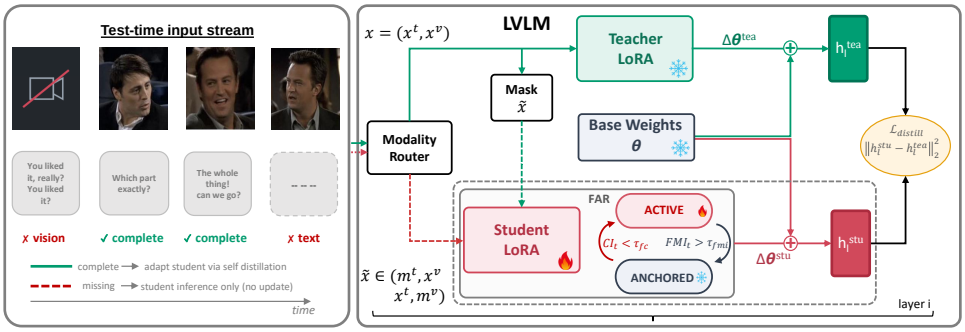

Figure 2: Illustration of the proposed TTSD-FAR method. The solid green arrow shows the flow of information in the case of complete modalities, and the dashed red arrow shows the flow of information in the missing modality case. The solid green arrow shows both modalities, and the single dashed green arrow denotes the input ($\tilde{x}$) with masked input. In the missing case, the model bypasses adaptation and only infers with the student LoRA. In the complete case, the model uses complete modality features from the teacher LoRA and masked modality features from the student LoRA, and distillation loss $\mathcal{L}_{\text{distill}}$ is calculated to match the student representation with the teacher representation.

representation; and (ii) in non-stationary streams, parameter drift can overwrite previously consolidated knowledge. We therefore augment TTSD with Fisher-Anchored Restoration (FAR), a two-state mechanism that determines *when* adaptation should proceed and *when* it should be suspended, based solely on the geometry of the Fisher importance landscape. See appendix A.1 for the theoretical properties and proofs of TTSD-FAR.

3.2 Architecture Overview

Our TTSD-FAR framework operates on a single base video-language model f^{θ^0} with two distinct LoRA adapters. Figure 2 illustrates the architecture. The architecture enables efficient mode switching:

$$\text{Teacher mode:} \quad \hat{y}^{\text{tea}} = f^{\theta^0+\Delta\theta^{\text{tea}}}(x^{\text{v}}, x^{\text{t}}) \tag{6}$$

$$\text{Student mode:} \quad \hat{y}^{\text{stu}} = f^{\theta^0+\Delta\theta^{\text{stu}}}(\tilde{x}^{\text{v}}, \tilde{x}^{\text{t}}) \tag{7}$$

3.3 TTA: Student LoRA

At test time, when encountering a sample x with complete modalities, we perform online adaptation of the student LoRA through feature-level self-distillation from the teacher. By directly matching intermediate representations, the student adapter is encouraged to capture the internal geometry of the teacher representation space rather than imitating its final decisions. This preserves the diversity and richness of the teacher features and mitigates mode collapse, which is particularly critical under missing-modality scenarios.

Given a layer $l \in \mathcal{L}$ of the LLM backbone, the hidden representations are:

$$h_l^{\text{tea}} = \text{LLM}_l^{\theta^0+\Delta\theta^{\text{tea}}}(E^{\text{v}}(x^{\text{v}}), E^{\text{t}}(x^{\text{t}})) \in \mathbb{R}^{s\times d} \tag{8}$$

$$h_l^{\text{stu}} = \text{LLM}_l^{\theta^0+\Delta\theta^{\text{stu}}}(E^{\text{v}}(\tilde{x}^{\text{v}}), E^{\text{t}}(\tilde{x}^{\text{t}})) \in \mathbb{R}^{s\times d} \tag{9}$$

where s is the sequence length (the sum of video tokens and text tokens) and d is the hidden dimension. We minimize the mean squared error between student and teacher representations as:

$$\mathcal{L}_{\text{distill}} = \sum_{l\in\mathcal{L}} \alpha_l \cdot \|h_l^{\text{stu}} - \text{sg}(h_l^{\text{tea}})\|_2^2 \tag{10}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation to prevent backpropagation into the frozen teacher, and α_l are layer-specific weights.

Teacher features are extracted once per sample. To ensure complete isolation, features are detached from the teacher network because no gradient calculation is required in the teacher LoRA. Teacher and student forward passes do not share computational graph nodes, preventing gradient contamination. Without isolation, shared normalization statistics or attention caches can cause interference.

Why Self-Distillation for Missing Modalities?
The teacher, trained on complete modalities, has learned rich cross-modal features that fuse information from both vision and text. When a modality is missing at test time, the student can be viewed as solving an inverse problem:

$$\Delta\theta^{\mathrm{stu}} = \arg\min_{\Delta\theta} \mathbb{E}_{\tilde{x}} \left\| \mathbf{h}_{\mathrm{stu}}(\tilde{x}) - \mathbf{h}_{\mathrm{tea}}(x) \right\|_2^2 . \tag{11}$$

3.4 FISHER-ANCHORED RESTORATION (FAR)

Continuous adaptation of ϕ is not harmless: once the student has found a solution to Eq. (11), additional gradient steps on the same distribution can introduce stochastic parameter drift, while distribution shifts in the test stream can move the student away from the consolidated solution. Monitoring L_{distill} directly is unreliable as a stopping signal since the loss can plateau not because the student has converged, but because the current batch of samples is uniformly hard. We therefore monitor the stability of the parameter-wise gradient-importance landscape rather than the loss magnitude itself. Symmetrically, drift should not be inferred from raw loss spikes, which can be triggered by hard samples in a stationary stream. Instead, FAR monitors whether incoming gradients engage parameter dimensions that were relatively unimportant during consolidation. Importantly, the resulting statistic is used as a gradient-geometry mismatch signal rather than as a definitive statistical test for distributional shift. FAR formalizes this intuition through two complementary gradient-importance statistics and a two-state mechanism: ACTIVE (the student LoRA is being updated) and ANCHORED (the student LoRA is frozen).

Online Fisher-Inspired Gradient-Importance Diagonal. After each adaptation step, we maintain an exponential moving average of the squared gradients of the distillation objective:

$$F_t = \beta_F F_{t-1} + (1 - \beta_F) \left(\nabla_\phi L_{\mathrm{distill}} \right)^{\odot 2} , \tag{12}$$

where $(\cdot)^{\odot 2}$ denotes element-wise squaring and $\beta_F \in (0, 1)$ is the decay parameter controlling the weight placed on historical estimates. We refer to F_t as an *online Fisher proxy*, since it captures the parameter-wise second moment of the gradients of the feature-level distillation objective rather than the classical Fisher information associated with a likelihood model. The element $F_{t,j}$ therefore provides a running estimate of the squared-gradient importance of the j-th parameter with respect to the distillation objective. F_t is updated only during the ACTIVE state.

Consolidation Index and the ACTIVE → ANCHORED Transition. The consolidation index (CI) measures the normalized ℓ_1 rate of change of the gradient-importance landscape between consecutive adaptation steps:

$$CI_t = \frac{\left\| F_t - F_{t-1} \right\|_1}{\left\| F_t \right\|_1 + \epsilon}, \tag{13}$$

where $\epsilon > 0$ is a small numerical stability constant. When CI_t is large, the gradient-importance landscape is still evolving. As the student settles into a stable adaptation regime, CI_t becomes small, indicating that the parameter-wise second-moment estimate has stabilized. The ℓ_1 norm is used because F_t is non-negative, and it measures the total mass shifted in the parameter-importance landscape.

The transition from ACTIVE to ANCHORED is triggered when CI_t remains below a threshold τ_{fc} for K consecutive adaptation steps. The student LoRA parameters ϕ are then frozen; no updates are applied in the ANCHORED state, and the current gradient-importance estimate is stored as the anchor, $F_{\mathrm{anchor}} = F_t$.

Fisher Mismatch Index and the ANCHORED → ACTIVE Transition. While ANCHORED, for each incoming complete-modality sample, the gradient of L_{distill} is computed at the frozen anchor parameters without applying a weight update:

$$g_t = \nabla_\phi L_{\text{distill}}|_{\phi=\phi_{\text{anchor}}}. \tag{14}$$

The Fisher mismatch index (FMI) for sample t is then defined as

$$FMI_t = \frac{1}{p}\sum_{j=1}^{p}\frac{g_{t,j}^2}{F_{\text{anchor},j}+\epsilon}, \tag{15}$$

where $p = |\phi|$ is the total number of student LoRA parameters. The ratio $g_{t,j}^2/(F_{\text{anchor},j}+\epsilon)$ becomes large when the current sample induces a strong gradient in a parameter direction that had relatively low gradient importance during consolidation. This provides a measure of mismatch between the current gradient geometry and the consolidated gradient-importance landscape. The transition from ANCHORED back to ACTIVE is triggered when $FMI_t > \tau_{fmi}$.

4 Results and Discussion

4.1 Experimental Methodology

Datasets. We evaluate our approach on three widely used multimodal and video-based emotion recognition benchmarks: Multimodal Emotion Lines Dataset **(MELD)** Poria et al. (2019), Dynamic Facial Expression in-the-Wild **(DFEW)** Jiang et al. (2020), and Behavioral Ambivalence/Hesitancy **(BAH)** González-González et al. (2026), each presenting distinct challenges in terms of modality diversity, temporal dynamics, and subtle affective cues. Details of the datasets are provided in the supplementary material. We exclusively validate on emotion recognition, which is one of the few multimodal applications where text transcripts serve as a crucial, non-redundant input modality. This makes it an ideal testbed for missing-modality robustness, since dropping the text modality induces genuine information loss.

Evaluation Protocol. This paper makes no assumptions about the training phase. TTSD-FAR only requires a model that works with multiple modalities at test time. To the best of our knowledge, we are among the first to study parameter-efficient test-time adaptation to missing modalities specifically in LVLMs. TTSD-FAR updates the model only when it encounters modality-complete samples but generates predictions for all samples, regardless of the modalities they contain. Missing modality scenarios are simulated by randomly masking modalities at inference time, with missing ratios ranging from 0% (complete) to 50%. 100% missing scenario is different from missing modalities, as it drifts into privileged modality literature and is outside the scope of this work. All baselines use the same frozen LVLM backbone for fair comparison.

4.2 Results under Missing Modalities

Table 1: Comparison of methods under progressively missing textual and visual modalities. We report F1 on the MELD dataset as the proportion of unavailable input increases. N/A: Perplexity Gen. is not applicable to the vision-missing case.

		Text Missing					**Vision Missing**				
Method	**Full**	**10%**	**20%**	**30%**	**40%**	**50%**	**10%**	**20%**	**30%**	**40%**	**50%**
No Adaptation	0.6149	0.5926	0.5706	0.5390	0.5190	0.4785	0.6122	0.6025	0.5934	0.5825	0.5775
TENT (ICLR'21)	–	0.1650	0.1680	0.1765	0.1855	0.1655	0.1658	0.1725	0.1655	0.1838	0.1826
EATA (ICML'22)	–	0.2854	0.2690	0.2770	0.2736	0.2482	0.2950	0.2750	0.2745	0.2530	0.2382
RAG (ICCV'25)	–	0.5640	0.5248	0.5120	0.4845	0.4660	0.5925	0.5830	0.5634	0.5567	0.5534
Perplexity Gen.	–	0.5680	0.5460	0.5255	0.4938	0.4722	N/A	N/A	N/A	N/A	N/A
TTSD-FAR (Ours)	–	**0.6115**	**0.5890**	**0.5625**	**0.5454**	**0.5124**	**0.6130**	**0.6075**	**0.6055**	**0.5970**	**0.5915**

Tables 1– 3 summarize performance across MELD, DFEW, and BAH under progressively increasing missing-modality ratios. Across all three datasets, entropy-based TTA methods (TENT and EATA)

Table 2: Comparison of methods under progressively missing visual modality. We report F1 on the DFEW dataset as the proportion of unavailable visual input increases.

Method	Full	10%	20%	30%	40%	50%
No Adap.	**0.5549**	0.5300	0.5078	0.4984	0.4837	0.4434
TENT	–	0.1548	0.1553	0.1665	0.1658	0.1640
EATA	–	0.2534	0.2778	0.2735	0.2845	0.2885
RAG	–	0.5340	0.5048	0.5020	0.4800	0.4468
(Ours)	–	**0.5515**	**0.5300**	**0.5195**	**0.5015**	**0.4775**

Table 3: Comparison of methods under progressively missing textual modality conditions for BAH dataset.

Method	Full	10%	20%	30%	40%	50%
No Adap.	**0.7142**	0.6835	0.6742	0.6526	0.6215	0.6023
TENT	–	0.4923	0.4834	0.4710	0.4572	0.4550
EATA	–	0.5630	0.5467	0.5360	0.5310	0.5250
RAG	–	0.6835	0.6730	0.6535	0.6215	0.5950
Perp. Gen.	–	0.6870	0.6745	0.6535	0.6310	0.6030
(Ours)	–	**0.7050**	**0.6925**	**0.6695**	**0.6405**	**0.6250**

consistently underperform, confirming that confidence-based objectives provide unreliable supervision for autoregressive LVLMs under severe modality shift. This observation is consistent with the findings of Hu et al. (2025), who showed that entropy minimization does not provide reliable optimization signals for autoregressive language models at test time. RAG and perplexity-based generation remain competitive when the observed modality is sufficiently informative, but their performance degrades as the dominant modality becomes unavailable.

On MELD (Table 1), removing text causes substantially greater degradation than removing vision ($0.6149 \rightarrow 0.4785$ versus $0.6149 \rightarrow 0.5775$), indicating that textual information provides the dominant supervision signal for emotion recognition in conversational settings. Consequently, retrieval-based methods struggle once textual information is unavailable, while entropy-based methods rapidly collapse toward near-random performance. TTSD-FAR instead leverages complete-modality teacher representations to supervise a student operating on synthetically masked inputs, allowing it to remain close to the full-modality baseline across both missing-text and missing-vision settings. A similar trend is observed on DFEW (Table 2), where removing visual information induces a substantial performance drop because facial dynamics constitute the dominant modality. Entropy-based adaptation again fails, while RAG deteriorates as weaker visual representations reduce retrieval quality. TTSD-FAR consistently achieves the strongest performance across all missing ratios, demonstrating that feature-level self-distillation provides a substantially more reliable adaptation signal than confidence- or retrieval-based approaches. Results for the weaker text-missing scenario are included in the supplementary material. Results on BAH (Table 3) further demonstrate the generality of the proposed framework. Although entropy-based methods perform relatively better because BAH is a binary classification task, they remain consistently inferior to TTSD-FAR. RAG and perplexity-based generation improve upon the zero-shot baseline at lower missing ratios but gradually deteriorate as textual information becomes increasingly unavailable. Additional experiments (Appendix B) and implementation details (Appendix D) are provided in the supplementary material.

4.3 Mechanistic Analysis of FAR

To isolate the contribution of FAR, we compare TTSD with and without the restoration mechanism across MELD and BAH under increasing missing rates. As shown in Figure 3, the unbounded variant (TTSD w/o FAR), which continues to adapt the student LoRA on every incoming sample, consistently underperforms the FAR-governed version at every missing rate. The consistent FAR gain across both the 7-class (MELD) and binary (BAH) settings indicates that this stopping/restarting behavior is not task-specific; rather, it reflects a general property of self-distillation under non-stationary missing-modality streams.

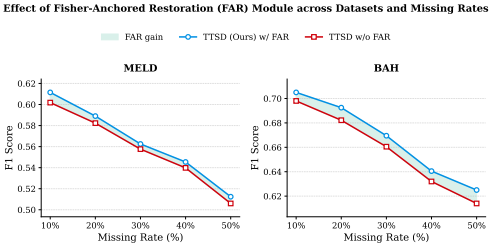

Figure 3: Effectiveness of the FAR module on the distillation results across different missing rates.

The three heatmaps in Figure 4 show how Fisher parameter importance is distributed across LoRA modules under two adaptation regimes. The left panel shows F_{anchor}, the Fisher diagonal stored at the moment FAR declared convergence, where importance is sharply concentrated in the early

layers of the query and up-projection modules, with the remainder of the landscape several orders of magnitude lower; this sparsity indicates that the student identified a precise, low-dimensional parameter subspace sufficient to solve the missing-modality distillation task. The middle panel shows the equivalent Fisher diagonal accumulated by continuous TTSD without any stopping criterion. The importance landscape is substantially more diffuse, and the gradient mass is spread uniformly across all layers and modules, reflecting that continued adaptation beyond convergence dilutes the signal that was meaningful at true convergence. The right panel shows the log-scale difference, which is almost entirely red, confirming that FAR anchors at a moment of higher and more concentrated importance across most of the parameter space. Additional ablations on FAR hyperparameters and state-transition visualizations are provided in Appendix C.2.

4.4 Discussion

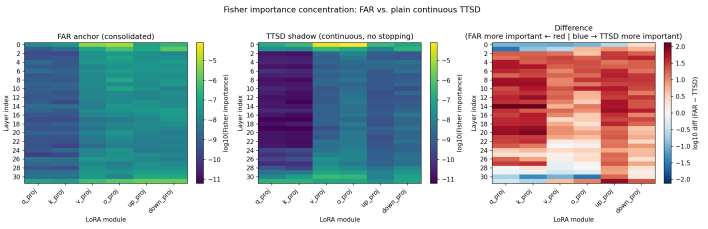

Figure 4: Fisher importance heatmaps ($\log_{10}$ scale): TTSD-FAR (left), TTSD (middle), and difference map (right), confirming systematically higher importance under FAR.

Our experiments reveal two core challenges for adapting autoregressive LVLMs under partial modality observation: estimating missing semantic content and sustaining effective adaptation over long deployment. Retrieval-based adaptation assumes the observed modality provides a query informative enough to retrieve semantically relevant references. This assumption breaks down when the missing modality carries the dominant information, in which case the retrieved examples supply inconsistent or misleading supervision. Entropy minimization fails in a complementary way. It implicitly assumes confident predictions are correct predictions, but under substantial modality shift, autoregressive LVLMs often generate overconfident, semantically incorrect outputs. Minimizing entropy in this regime reinforces these errors, producing optimization drift rather than genuine adaptation, consistent with the findings of Hu et al. (2025).

TTSD-FAR addresses both failure modes by treating missing-modality adaptation as a conditional representation alignment problem rather than a confidence or retrieval optimization problem. The student is guided toward the frozen reference model's complete-modality latent representation through feature-level supervision, estimating the complete-modality representation conditioned on the available modalities and yielding substantially more robust adaptation under severe degradation. A separate issue is that continued optimization past convergence causes parameter diffusion, eroding useful representations. FAR's adaptive consolidation addresses this by tracking optimization stationarity via Fisher geometry to detect convergence, then reactivating adaptation when sufficient mismatch with the consolidated gradient-importance geometry is observed, which we take as evidence of likely distributional drift.

5 Conclusion

TTSD-FAR is a parameter-efficient framework for adapting large video-language models to missing modalities during inference. Missing modalities induce substantial changes in representation geometry within autoregressive LVLMs, a regime in which entropy-based test-time adaptation and retrieval-augmented generation can provide unreliable supervision. TTSD-FAR instead couples a frozen teacher, trained on complete modalities, with an adaptive student adapter that operates on masked inputs, aligning student and teacher representations via feature-level self-distillation without modifying the base model's parameters. Fisher-Anchored Restoration governs this process through a cycle of consolidation and reactivation: adaptation is suspended once the student's parameter importance landscape stabilizes, and resumed when sufficient mismatch with the consolidated gradient-importance geometry is observed, reducing unnecessary parameter drift while retaining the capacity to adapt to changing test conditions. Extensive experiments across three emotion recognition benchmarks show that TTSD-FAR improves performance under severe missing-modality conditions while maintaining performance close to the complete-modality setting.

AI USE STATEMENT

In this work, we used generative AI tools for improving the writing and language of the manuscript, and for LaTeX formatting assistance (e.g., table formatting and structuring). We have not used generative AI tools for research ideation, experimental design, or generating research hypotheses. Tasks such as synthetic data generation and data analysis do not apply to this work. Additionally, we used generative AI tools for polishing the phrasing of explanatory text and for suggesting section organization. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on improving the robustness of large vision-language models (LVLMs) to missing modalities through test-time self-distillation, and is conducted entirely on publicly available benchmark datasets widely used in the multimodal learning community; we do not collect, annotate, or release any new data involving human subjects. We believe the proposed method has a positive societal impact by making multimodal systems more reliable under real-world conditions such as missing sensor streams, and we do not foresee direct negative ethical consequences arising from this work. We have no conflicts of interest to disclose. All datasets and pretrained models used are properly cited and used in accordance with their respective licenses.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure that the results presented in this paper are reproducible. The proposed TTSD-FAR method is described in detail in Section 3, including the test-time self-distillation procedure and the missing-modality simulation protocol used to generate the missing rates evaluated in our experiments. Implementation details, including model architectures, hyperparameters, and training/optimization settings, are provided in the Appendices. All experiments are conducted on publicly available datasets, and details of the data preprocessing and missing-modality generation procedure are included in the Appendix. For experiments where different seeds are used, the exact seed value is presented in the caption. We will release our source code (also provided in the supplementary material) upon publication to facilitate independent verification and reproduction of our results.

REFERENCES

Mohammad Mahdi Abootorabi et al. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In *Findings of ACL*, 2025.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *ICLR*, abs/2310.11511, 2023. URL https://api.semanticscholar.org/CorpusID:264288947.

Junsong Chen, Jiyuan Liu, Suyuan Liu, Wei Zhang, Ao Li, En Zhu, and Xinwang Liu. Sample-specific modality diagnosis and cross-modal enhancement for incomplete multimodal representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(24):20154–20162, Mar. 2026. doi: 10.1609/aaai.v40i24.39102. URL https://ojs.aaai.org/index.php/AAAI/article/view/39102.

Mengying Ge, Dongkai Tang, and Mingyang Li. Video emotion open-vocabulary recognition based on multimodal large language model, 2024.

M. González-González, S. Belharbi, M. O. Zeeshan, M. Sharafi, M. H Aslam, M. Pedersoli, A. L. Koerich, S. L. Bacon, and E. Granger. BAH dataset for ambivalence/hesitancy recognition in videos for digital behavioural change. In *ICLR*, 2026.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proc. ACL*, pp. 1726–1736, 2024.

Jinwu Hu, Zhitian Zhang, Guohao Chen, Xutao Wen, Chao Shuai, Wei Luo, Bin Xiao, Yuanqing Li, and Mingkui Tan. Test-time learning for large language models, 2025.

Dawei Huang, Qing Li, Chuan Yan, Zebang Cheng, Zihao Han, Yurong Huang, Xiang Li, Bin Li, Xiaohui Wang, Zheng Lian, Zhi-Qi Cheng, and Xiaojiang Peng. Emotion-qwen: A unified framework for emotion and vision understanding, 2025. URL https://arxiv.org/abs/2505.06685.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild, 2020.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526, March 2017. ISSN 1091-6490. doi: 10.1073/pnas.1611835114. URL http://dx.doi.org/10.1073/pnas.1611835114.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *arXiv preprint arXiv:2311.10122*, 2023.

Guan-Ting Lin, Wei Ping Huang, and Hung-yi Lee. Continual test-time adaptation for end-to-end speech recognition on noisy speech. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

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

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.679.

Harsh Maheshwari, Yen-Cheng Liu, and Zsolt Kira. Missing modality robustness in semi-supervised multi-modal semantic segmentation. In *Proc. IEEE/CVF WACV*, pp. 1020–1030, 2024.

Himanshu Mishra. RDumb++: Drift-aware continual test-time adaptation. *arXiv preprint arXiv:2601.15544*, 2026.

Zachary Nado et al. Evaluating prediction-time batch normalization for robustness under covariate shift. *NeurIPS*, 2021.

Natalia Neverova et al. Moddrop: Adaptive multi-modal gesture recognition. In *CVPR*, 2016.

Niki Nezakati, Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and Salman Asif. Mmp: Towards robust multi-modal learning with masked modality projection, 10 2024.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 17–23 Jul 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world, 2023. URL https://arxiv.org/abs/2302.12400.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations, 2019.

Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge. RDumb: A simple approach that questions our progress in continual test-time adaptation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 39915–39935, 2023.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 2366–2377, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714805.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, pp. 8748–8763. PMLR, 2021.

Merey Ramazanova, Alejandro Pardo, Bernard Ghanem, and Motasem Alfarra. Test-time adaptation for combating missing modalities in egocentric videos, 2025.

Muhammad Saad Saeed, Shah Nawaz, Muhammad Zaigham Zaheer, Muhammad Haris Khan, Karthik Nandakumar, Muhammad Haroon Yousaf, Hassan Sajjad, Tom De Schepper, and Markus Schedl. Modality invariant multimodal learning to handle missing modalities: A single-branch approach, 2024.

Abdelrahman Shaker, Muhammad Maaz, Chenhui Gou, Hamid Rezatofighi, Salman Khan, and Fahad Shahbaz Khan. Mobile-videogpt: Fast and accurate video understanding language model, 2025. URL https://arxiv.org/abs/2503.21782.

Yu Sun et al. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020.

Dequan Wang et al. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021.

Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1828–1838, 2020.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, 2022.

Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. LongRAG: A dual-perspective retrieval-augmented generation paradigm for long-context question answering. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22600–22632, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1259.

SUPPLEMENTARY MATERIAL

CONTENTS

Algorithm for TTSD-FAR 14

Theoretical Properties of FAR 14

Proposition 1: Fisher Stability as a Stationarity Signal 15

Proposition 2: Conditional Reactivation Under Fisher Mismatch 16

Proposition 3: High-loss Hard Samples do not necessarily trigger FMI 16

Additional Results 17

Distilling from Base Teacher (no teacher LoRA) 17

Weaker Modality Missing Results 17

Additional Ablations 18

LoRA Rank Sensitivity 18

Effectivity of TTSD on Small Video Language Model 19

FAR Hyperparameters Sensitivity 20

Effect of the FMI Threshold 20

FAR State Dynamics over the Test Stream 20

Loss-Based Stopping vs. FAR 22

Datasets and Implementation Details 21

Datasets 21

Implementation Details 23

FAR Parameters 24

A ALGORITHM FOR TTSD-FAR

Algorithm 1 summarizes the adaptation loop for the LVLM using the proposed TTSD-FAR methodology.

Algorithm 1 TTSD-FAR

Require: $\Delta\theta^{\text{tea}}$, ϕ_0, η, β_F, τ_{fc}, M, K, τ_{fmi}, ε
$\phi \leftarrow \phi_0$; $F, F_{\text{prev}} \leftarrow \mathbf{0}_p$; state $\leftarrow$ ACTIVE; $c \leftarrow 0$; $\mathcal{B} \leftarrow \emptyset$
for each incoming sample $\tilde{x}$ **do**
  if state $=$ ACTIVE **then**
    $h_l^{\text{tea}} \leftarrow \text{sg}(\text{LLM}_l^{\theta^0+\Delta\theta^{\text{tea}}}(x))$; $h_l^{\text{stu}} \leftarrow \text{LLM}_l^{\theta^0+\Delta\theta^{\text{stu}}}(\tilde{x})$
    $g \leftarrow \nabla_\phi \sum_l \alpha_l \|h_l^{\text{stu}} - h_l^{\text{tea}}\|_2^2$; $\phi \leftarrow \phi - \eta g$
    $F \leftarrow \beta_F F_{\text{prev}} + (1-\beta_F) g^{\odot 2}$
    $\text{CI} \leftarrow \|F - F_{\text{prev}}\|_1 / (\|F\|_1 + \varepsilon)$; $F_{\text{prev}} \leftarrow F$
    $c \leftarrow c+1$ if $\text{CI} < \tau_{\text{fc}}$, else $c \leftarrow 0$
    if $c \geq K$ **then**
      $\phi_{\text{anchor}}, F_{\text{anchor}} \leftarrow \phi, F$; state $\leftarrow$ ANCHORED; $\mathcal{B} \leftarrow \emptyset$
    end if
  else
    Infer with frozen ϕ_{anchor} {no update}
    if $\tilde{x}$ is a complete-modality sample **then**
      Mask $\tilde{x}$ synthetically $\rightarrow \tilde{x}'$; $g \leftarrow \nabla_\phi \mathcal{L}_{\text{distill}}(\tilde{x}')|_{\phi_{\text{anchor}}}$
      $\text{FMI} \leftarrow \frac{1}{p}\sum_j g_j^2 / (F_{\text{anchor},j} + \varepsilon)$; push to $\mathcal{B}$ (keep last M)
      if $\text{FMI} > \tau_{fmi}$ **then**
        $\phi, F_{\text{prev}} \leftarrow \phi_{\text{anchor}}, F_{\text{anchor}}$; $c \leftarrow 0$; state $\leftarrow$ ACTIVE; $\mathcal{B} \leftarrow \emptyset$
      end if
    end if
  end if
end for

A.1 THEORETICAL PROPERTIES OF FAR

This section provides an intuitive theoretical justification for the proposed Fisher-Anchored Restoration (FAR) mechanism. Rather than proving global optimality, our goal is to explain why the proposed two-state adaptation strategy avoids unnecessary parameter drift while remaining capable of re-adapting under distribution shift.

A.1.1 ASSUMPTIONS

The following assumptions are standard in stochastic optimization and test-time adaptation.

Assumption 1 (Local Smoothness). The feature-level distillation objective $L_{\text{distill}}(\phi)$ is continuously differentiable with L-Lipschitz gradients,

$$\|\nabla L(\phi_1) - \nabla L(\phi_2)\| \leq L\|\phi_1 - \phi_2\|. \tag{16}$$

Assumption 2 (Bounded Gradient Variance). The stochastic gradients satisfy

$$\mathbb{E}[g_t] = \nabla L(\phi_t), \tag{17}$$

and

$$\mathbb{E}\left[\|g_t - \nabla L(\phi_t)\|^2\right] \leq \sigma^2. \tag{18}$$

Assumption 3 (Piecewise Stationary Test Stream). The test distribution remains stationary between distribution shifts. This assumption is common in continual test-time adaptation, where environmental conditions remain approximately constant for finite intervals before changing.

A.1.2 Proposition 1: Fisher Stability as a Stationarity Signal.

Let F_t denote the online Fisher diagonal defined in Eq. (12),

$$F_t = \beta_F F_{t-1} + (1-\beta_F)\, g_t^{\odot 2}, \tag{19}$$

where g_t is the stochastic gradient of $\mathcal{L}_{\text{distill}}$ at step t, taken coordinate-wise so that $g_t^{\odot 2}$ denotes element-wise squaring. Under Assumption 2 (bounded gradient variance) and Assumption 3 (piece-wise stationarity), suppose

$$\mathrm{CI}_t = \frac{\|F_t - F_{t-1}\|_1}{\|F_t\|_1 + \varepsilon} < \tau_{fc} \tag{20}$$

holds for K consecutive adaptation steps. Then the second moment of the stochastic gradient, $\mathbb{E}[g_t^{\odot 2}]$, has stabilized across those steps, in the sense of Eq. (25) below. This need not imply that $\nabla\mathcal{L}(\phi_t) \approx 0$.

Proof Sketch: From the recursion defining F_t,

$$F_t - F_{t-1} = (1-\beta_F)\left(g_t^{\odot 2} - F_{t-1}\right). \tag{21}$$

Substituting into the definition of CI_t,

$$\mathrm{CI}_t = \frac{(1-\beta_F)\,\|g_t^{\odot 2} - F_{t-1}\|_1}{\|F_t\|_1 + \varepsilon}. \tag{22}$$

Since $\beta_F \in (0,1)$ is fixed, $\mathrm{CI}_t < \tau_{fc}$ is equivalent, up to the constant $(1-\beta_F)$, to

$$\|g_t^{\odot 2} - F_{t-1}\|_1 < \frac{\tau_{fc}}{1-\beta_F}\left(\|F_t\|_1 + \varepsilon\right), \tag{23}$$

i.e., the current squared gradient agrees with the running Fisher estimate in aggregate ℓ_1 mass. Taking expectations under Assumption 2, which gives $\mathbb{E}[g_t] = \nabla\mathcal{L}(\phi_t)$ and $\mathbb{E}\|g_t - \nabla\mathcal{L}(\phi_t)\|^2 \le \sigma^2$, the per-coordinate second moment satisfies

$$\mathbb{E}[(g_{t,j})^2] = (\nabla_j \mathcal{L}(\phi_t))^2 + \mathrm{Var}(g_{t,j}). \tag{24}$$

Condition (20) small for K consecutive steps therefore forces

$$\mathbb{E}[(g_{t,j})^2] \approx \mathbb{E}[(g_{t-1,j})^2] \approx \cdots \approx \mathbb{E}[(g_{t-K,j})^2] \quad \forall j, \tag{25}$$

which holds under either of two disjoint cases:

$$\text{(a) stationary point:} \quad \nabla\mathcal{L}(\phi_t) \approx 0, \tag{26}$$

$$\mathrm{Var}(g_{t,j}) \approx \sigma_j^2, \tag{27}$$

$$\text{(b) constant residual:} \quad \nabla\mathcal{L}(\phi_t) \approx c_j \neq 0, \tag{28}$$

$$\mathrm{Var}(g_{t,j}) \approx \sigma_j^2 - c_j^2 \quad \text{(unchanged)}. \tag{29}$$

Equation (25) alone cannot distinguish case (a) from case (b): both produce a stable F_t. What Eq. (25) does establish is that the distribution generating g_t, and in particular its second moment, has stopped changing over the K-step window.

Remark (Informal): Consolidation Reduces Unnecessary Drift. Given Eq. (25), consider T additional adaptation steps while $CI_t < \tau_{fc}$. Decompose the stochastic gradient as

$$g_t = \nabla L(\phi_t) + \xi_t, \tag{30}$$

where ξ_t denotes the stochastic gradient noise. Under the additional standard assumption that $\{\xi_t\}$ forms a martingale-difference sequence with bounded variance,

$$\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}\left[\|\xi_t\|_2^2\right] \le \sigma^2, \tag{31}$$

the variance of the accumulated stochastic component of the parameter update satisfies

$$\mathrm{Var}\left(\eta \sum_{i=1}^{T} \xi_{t+i}\right) = O(T\eta^2\sigma^2). \tag{32}$$

This result does not imply that the total parameter displacement is $O(T)$, since a nonzero mean gradient can introduce an additional systematic drift term. Rather, it shows that continued adaptation can accumulate stochastic update variance even after the gradient second-moment landscape has stabilized.

FAR therefore freezes ϕ once the consolidation criterion holds for K consecutive steps, trading potentially small residual optimization progress for reduced stochastic parameter drift until the FMI criterion indicates that reactivation may be warranted.

A.1.3 PROPOSITION 2: CONDITIONAL REACTIVATION UNDER FISHER MISMATCH

Assume the student is in the ANCHORED state and the test distribution changes from P_A to P_B. If the resulting gradient statistics satisfy

$$FMI_t > \tau_{fmi}, \tag{33}$$

then FAR exits the ANCHORED state and resumes gradient updates. Following reactivation, the optimization update is identical to continuous TTSD adaptation, up to the finite detection delay.

Proof Sketch. During the ANCHORED state, the student parameters remain fixed:

$$\phi_t = \phi_{\text{anchor}}. \tag{34}$$

For each incoming complete-modality sample, FAR evaluates

$$FMI_t = \frac{1}{p}\sum_{j=1}^{p} \frac{g_{t,j}^2}{F_{\text{anchor},j} + \epsilon}. \tag{35}$$

By the transition rule, whenever $FMI_t > \tau_{fmi}$, FAR exits the ANCHORED state and reactivates adaptation.

After reactivation, FAR performs the same parameter update as standard TTSD:

$$\phi_{t+1} = \phi_t - \eta g_t. \tag{36}$$

Therefore, conditional on the FMI threshold being crossed, FAR resumes the same optimization dynamics as continuous TTSD after the detection delay. The proposition does not require that every distributional shift cross the FMI threshold; rather, it establishes the behavior of FAR once sufficient gradient-geometry mismatch is observed.

A.1.4 PROPOSITION 3: HIGH-LOSS HARD SAMPLES DO NOT NECESSARILY TRIGGER FMI.

Suppose the student LoRA is in the ANCHORED state with stored baseline

$$F_{\text{anchor}} \approx F^*(P) := \mathbb{E}_{x\sim P}\left[g(x) \odot g(x)\right]. \tag{37}$$

Consider a sample x_H drawn from the same stationary distribution P but having unusually high distillation loss. If its expected squared-gradient geometry remains consistent with the anchored gradient-importance landscape, i.e.,

$$\mathbb{E}_{x_H\sim P}\left[g_H \odot g_H\right] \approx F^*(P) \approx F_{\text{anchor}}, \tag{38}$$

then

$$\mathbb{E}_{x_H\sim P}\left[FMI(x_H)\right] \approx \frac{1}{p}\sum_{j=1}^{p} \frac{F_{\text{anchor},j}}{F_{\text{anchor},j} + \epsilon} \approx 1. \tag{39}$$

Thus, elevated distillation loss by itself does not imply a large FMI. In contrast, a distributional change that shifts gradient mass toward parameter directions that had low importance during consolidation can increase the FMI and potentially trigger reactivation.

Proof Sketch. By definition, the stationary gradient-importance landscape is

$$F^*(P)_j = \mathbb{E}_{x\sim P}\left[g_j(x)^2\right]. \tag{40}$$

For a hard sample drawn from the same stationary distribution, high distillation loss does not necessarily imply a change in the expected squared-gradient geometry. Under Eq. 38,

$$\mathbb{E}\left[g_{H,j}^2\right] \approx F_{\text{anchor},j}. \tag{41}$$

Therefore,

$$\mathbb{E}\left[FMI(x_H)\right] \approx \frac{1}{p}\sum_{j=1}^{p} \frac{F_{\text{anchor},j}}{F_{\text{anchor},j} + \epsilon} \approx 1. \tag{42}$$

By contrast, when a shifted distribution produces relatively larger gradients in coordinates with small $F_{\text{anchor},j}$, the corresponding ratios $g_j^2/(F_{\text{anchor},j}+\epsilon)$ increase, resulting in a larger FMI. Importantly, this establishes FMI as a gradient-geometry mismatch signal rather than a definitive statistical test for distributional shift.

Table 4: Comparison of methods under progressively missing text conditions. Distilling from the Base model, i.e., no teacher LoRA. We report F1 score on the BAH dataset as the proportion of unavailable textual input increases.

Method	Full	10%	20%	30%	40%	50%
No Adaptation	**0.6341**	0.5905	0.5798	0.5710	0.5662	0.5614
TENT	–	0.4910	0.4840	0.4733	0.4501	0.4453
EATA	–	0.5465	0.5367	0.5290	0.5245	0.5110
RAG	–	0.5770	0.5720	0.5540	0.5448	0.5250
Perplexity Gen.	–	0.5790	0.5695	0.5648	0.5488	0.5430
TTSD-FAR (Ours)	–	**0.6185**	**0.6120**	**0.5955**	**0.5910**	**0.5870**

B Additional Results

Distilling from Base Teacher (no teacher LoRA).

When working with LLMs and LVLMs, it is often realistic to assume that no retraining, either full or parameter-efficient, is performed on the base model. To demonstrate the effectiveness of TTSD under this constraint, we conduct an experiment where distillation is performed directly from the base model weights without any teacher LoRA as shown in Fig. 5. In this setting, teacher representations are extracted from the frozen base model, resulting in the same optimization problem as Eq. 11

Table 4 shows the results for the case where the teacher embeddings are obtained directly from the base model weights. Since the same distillation mechanism is applied, the setup is effectively identical except that the teacher representations are extracted from the frozen base model rather than from a Teacher LoRA trained with the objective in Eq. 11. The results therefore follow a similar trend to the case where the teacher embeddings are obtained using the Teacher LoRA objective.

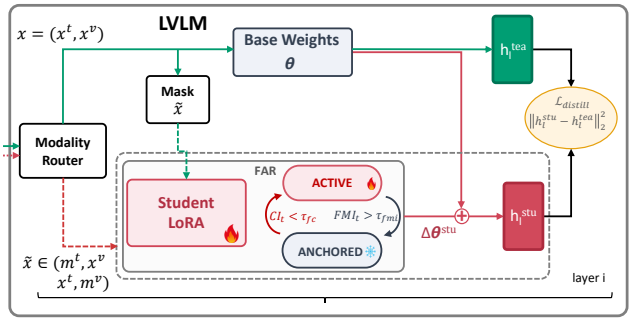

Figure 5: TTSD-FAR without teacher LoRA. Base weights provide teacher representations.

Weaker-Modality Missing Results

DFEW - Text Missing at Test-Time - Weaker Modality Missing

Table 5 reports the performance of different methods on the DFEW under progressively increasing levels of missing textual modality. As the proportion of missing text increases, the performance of the model without adaptation gradually degrades.

However, the performance drop due to the missing modality is insignificant compared to the vision missing in the DFEW dataset. This is primarily due to the dataset structure. The DFEW dataset comprises of 16372 short clips, of which 6196 are non-English transcriptions and 4960 clip have no speaker utterance at all. This shows that even the full modality performance (0.5549) is already less reliant on the textual modality. Classical test-time adaptation methods such as TENT Wang et al. (2021) and EATA Niu et al. (2022) exhibit severe performance drops, suggesting that entropy-minimization-based adaptation strategies are not well suited for multimodal missing-modality scenarios. In contrast, retrieval-based augmentation (RAG) and generation-based approaches main-

Table 5: Comparison of methods under progressively missing textual modality conditions. F1 on the DFEW dataset as the proportion of missing textual input increases.

Method	Full	10%	20%	30%	40%	50%
No Adaptation	**0.5549**	0.5530	0.5490	0.5365	0.5260	0.5235
TENT	–	0.1655	0.1725	0.1790	0.1935	0.1833
EATA	–	0.2635	0.2674	0.2835	0.2855	0.2943
RAG	–	0.5530	0.5425	0.5265	0.5265	0.5210
Perplexity Gen.	–	0.5535	0.5460	0.5350	0.5300	0.5210
TTSD-FAR (Ours)	–	**0.5540**	**0.5510**	**0.5444**	**0.5376**	**0.5278**

tain performance because the observed modality is stronger in this case. Our proposed **TTSD** method consistently achieves the best performance under moderate to severe missing-text conditions (20%–50%), demonstrating improved robustness and stability when textual inputs are partially unavailable. These results indicate that TTSD effectively mitigates the negative impact of missing textual modality while preserving strong performance as modality degradation increases.

BAH - Vision Missing at Test-Time- Weaker Modality Missing

Table 6 reports results on BAH under vision missing conditions. The vision drop for the BAH dataset results in negligible performance drops even at severe missing percentages (50%). The 10% has no effect on the performance. TTSD is able to improve over the existing TTA methods in both weaker and stronger missing modality conditions.

Table 6: Comparison of methods under progressively missing visual modality. We report F1 score on the BAH dataset as the proportion of unavailable visual input increases.

Method	Full	10%	20%	30%	40%	50%
No Adaptation	**0.7142**	0.7142	0.7095	0.7046	0.6950	0.6923
TENT	–	0.4834	0.4846	0.4790	0.4645	0.4754
EATA	–	0.5655	0.5428	0.5165	0.5095	0.5052
RAG	–	0.7100	0.7055	0.7059	0.7110	0.6975
TTSD-FAR (Ours)	–	**0.7142**	**0.7125**	**0.7120**	**0.7125**	**0.7125**

As shown in Table 9, the textual input is critical for Ambivalence detection. This explains why the text-missing case for the BAH dataset results in a much higher performance drop than vision-missing, as shown in Table 6. Another important observation is that in the vision missing case (Table 6) the RAG-based method is also able to improve, unlike the text missing case. This is because in the vision missing case, the observed modality (text) is stronger and is able to effectively retrieve relevant samples. Whereas, in the text missing case, the observed modality (vision) was not discriminant enough for effective retrieval.

C Additional Ablations:

LoRA Rank Sensitivity.

We investigate the effect of LoRA rank on TTSD's robustness to modality-incomplete inputs. As expected, performance degrades monotonically with increasing missing rates across all ranks; yet, the degradation pattern reveals a clear sensitivity to rank capacity. Lower-rank adaptation (r=4) suffers disproportionately under higher missing rates, exhibiting steeper and less stable declines compared to higher rank, suggesting that insufficient parameter capacity limits the model's ability to compensate for missing modalities. Ranks r=8 and r=16 remain consistently close throughout, with only marginal gains from doubling the rank, indicating diminishing returns beyond r=8. These results suggest that r=8 strikes the best balance between adaptability, robustness, and computational complexity. The ablation shows that LoRA rank is a meaningful factor in missing-modality robustness.

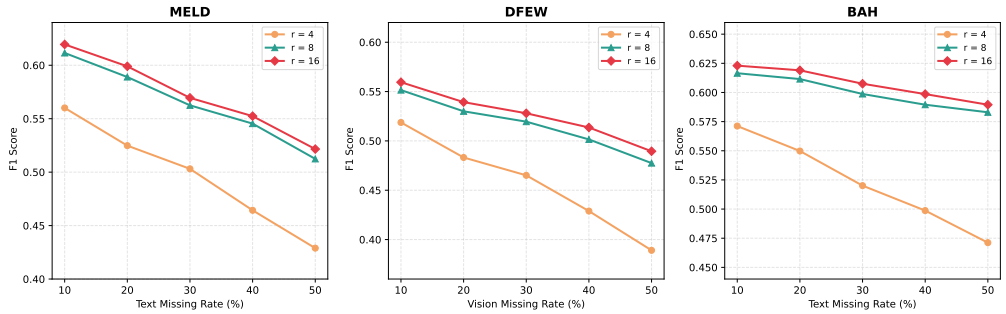

Figure 6: Impact of LoRA rank on distillation performance under varying missing-modality rates across three datasets.

Effectiveness and Stability of TTSD on a Small Video-Language Model (SVLM):

To further show the effectiveness of the proposed TTSD method, we perform the experiment on a smaller video language model, MobileVideoGPT-0.5B Shaker et al. (2025). We repeat each configuration across three random seeds (42, 1000, 1999) to assess whether the observed gains are consistent or an artifact of a particular run. We use this smaller model for the multi-seed analysis primarily for computational efficiency, since running three seeds across five missing-modality rates on a 7B-scale model would be considerably more expensive; the 0.5B model allows us to obtain a reliable estimate of variance at a fraction of the compute cost while still verifying that the underlying mechanism holds.

Table 7: Robustness to missing modalities across missing rates (mean ± SD across 3 seeds: 42, 1000, 1999; MobileVideoGPT-0.5B). Full-modality accuracy is 0.4889 for No Adaptation, constant across missing rates.

Method	0.1	0.2	0.3	0.4	0.5
No Adaptation	0.4519 ± 0.004	0.4421 ± 0.01	0.4306 ± 0.01	0.4220 ± 0.02	0.4127 ± 0.02
TTSD-FAR (Ours)	$\mathbf{0.4687 \pm 0.005}$	$\mathbf{0.4586 \pm 0.012}$	$\mathbf{0.4515 \pm 0.013}$	$\mathbf{0.4423 \pm 0.025}$	$\mathbf{0.4363 \pm 0.025}$

Applying TTSD consistently improves the results across all missing-modality levels, demonstrating that the proposed method effectively recovers useful representations when textual information becomes partially unavailable. Overall performance is lower than the Video-LLaVA-7B model due to its reduced model capacity, but TTSD still provides consistent improvements over the non-adapted baseline under all missing-modality conditions. These results indicate that TTSD is not restricted to a specific LVLM architecture and can improve robustness for both large and lightweight video-language models.

We further observe a consistent trend between missing rate and variance: standard deviations increase monotonically from the lowest missing rate (0.1) to the highest (0.5) for both methods. This is expected, since at low missing rates the model still has access to most of the original modality information, making its predictions largely insensitive to which specific seed is used. As the missing rate increases, a larger and more arbitrary portion of the input is dropped, so the remaining samples available to the model become more sensitive to the particular pattern of missingness sampled under each seed. Consequently, run-to-run variability increases with the severity of the missing-modality condition, for both the baseline and TTSD-FAR.

C.1 Computational Complexity Analysis

Experiments show that TTSD-FAR performs well when adapting to missing modality conditions at test time. However, this performance gain comes at a computational cost. We distinguish the two cases and explain the computational overhead for both cases. i) **Full modality case:** TTSD-FAR requires two forward passes, one forward pass for obtaining the teacher features and one for

obtaining the student features. ii) **Missing modality case:** For the missing modality case, there is no computational overhead. The model uses the adapted student LoRA for prediction. We update only 44M parameters out of the 7B parameters of the *Video-LLaVA 7B* model. This results in only $\approx 0.629\%$ of the parameter updates. It is also important to note that the total wall time for the adaptation scenario depends on the missing ratio. For the 50% missing case, 50% of the samples will pose no additional overhead; the remaining 50% of the input samples require two forward passes and backpropagation to the student LoRA. Consequently, the total wall time of the complete MELD test set without adaptation takes ≈ 1.5 hours, and the time taken with adaptation is ≈ 2.1 hours.

C.2 FAR Hyperparameters Sensitivity

Effect of the FMI Threshold:

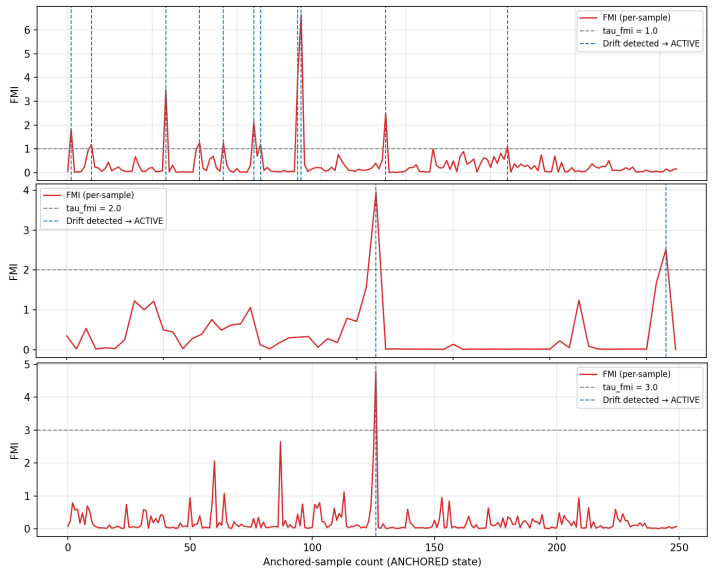

Figure 7: Effect of τ_{fmi} on reactivation events. The three panels show the FMI trace at τ_{fmi} 1.0 (top), 2.0 (middle) and, 3.0 (bottom).

Figure 7 illustrates the sensitivity of FAR's reactivation criterion to the choice of τ_{fmi}. Each panel plots the FMI trace computed during the ANCHORED state, with the horizontal dashed line indicating τ_{fmi} and vertical dashed lines marking each detected reactivation event. At $\tau_{fmi} = 1.0$ (top), the threshold is low enough that it is crossed both by the largest, genuine spikes and by several smaller fluctuations in the 1.0–1.5 range, yielding nine separate reactivations. This demonstrates the failure mode of an overly permissive threshold, where the mechanism reacts to noise as frequently as to true distributional shift. At $\tau_{fmi} = 2.0$ (middle), corresponding to the value used in our reported experiments, the threshold clears the minor fluctuations, while still detecting the two clearly dominant spikes, resulting in exactly two reactivations. This setting represents the balance adopted for the main results. At $\tau_{fmi} = 3.0$ (bottom), the threshold is raised further such that even a second sizeable spike is no longer detected, leaving only the single largest event to trigger reactivation. This illustrates the opposite failure mode: an excessively conservative threshold trades false positives for false negatives, risking missed detections of genuine distributional shift.

FAR State Dynamics over the Test Stream

Figure 8 shows the ACTIVE/ANCHORED state trajectory of FAR across the test stream under three hyperparameter regimes. The middle panel corresponds to the values reported in the paper and supplementary ($\beta_F = 0.99$, $\tau_{fc} = 0.02$, $K = 5$, $\tau_{fmi} = 2.0$). The student remains ACTIVE for roughly the first 40% of the stream before the Consolidation Index stays below τ_{fc} for K consecutive samples, triggering consolidation into ANCHORED. It remains frozen until a distributional shift shortly after, around the mid-point of the stream, raises the FMI above τ_{fmi}, triggering reactivation.

A second cycle follows later in the stream, consolidating and then reactivating again, after which the student remains active for the remainder until the very end, where it transitions to the Anchored state. This produces two well-separated reactivation cycles.

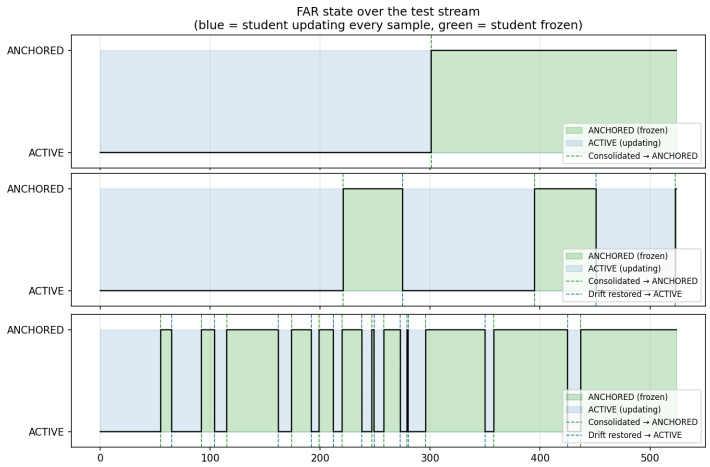

Figure 8: Timeline of FAR states (ACTIVE-ANCHORED) across the test stream.

The top panel shows a single consolidation slightly past the mid-point of the stream, with no reactivation for the remainder. This is consistent with a substantially higher $\tau_{fmi} = 5$, under which later shifts in the stream no longer raise the FMI enough to cross the threshold, leaving the student permanently frozen once anchored. The bottom panel uses a relaxed consolidation requirement with a lower reactivation threshold of $\tau_{fmi} = 0.5$. Consolidation fires as soon as CI drops below τ_{fc}, and the low τ_{fmi} causes ordinary gradient noise while anchored to be sufficient to trigger reactivation. These settings produce rapid, near-continuous alternation between ACTIVE and ANCHORED throughout most of the stream, resulting in the student being in the ANCHORED state for roughly 80% of the test stream. These two extremes motivate the reported configuration of $K = 5$ and $\tau_{fmi} = 2.0$, which yields the stable, interpretable cycles shown in the middle panel rather than premature permanent freezing or unstable oscillation.

C.3 Loss-Based Stopping vs. FAR

Table 8 compares TTSD-FAR against two ablated variants on the BAH dataset (with settings similar to Tab. 4) under progressively missing textual input. i) plain TTSD, which adapts on every complete-modality sample without any stopping criterion, and ii) TTSD-LS, which instead halts adaptation once the raw distillation loss plateaus with a patience of 5, after which the student remains frozen for the rest of the stream. TTSD-LS underperforms not only TTSD-FAR but also plain, unbounded TTSD at every missing ratio. This confirms the motivation given in Section 3.4: the raw distillation loss is an unreliable stopping signal, since it can plateau due to sample difficulty rather than genuine parameter convergence. Once frozen, TTSD-LS has no mechanism to detect that the test stream has since shifted and never resumes adaptation. It accumulates the disadvantages of stopping without any of the drift-awareness that FAR provides. In contrast, TTSD-FAR consistently outperforms both plain TTSD and TTSD-LS across all missing ratios, indicating that gating adaptation on the Fisher information geometry rather than on the loss itself is necessary to obtain the benefits of stopping without sacrificing the ability to resume adaptation when genuine distributional shift occurs.

D Datasets and Implementation Details

Datasets

MELD (Multimodal EmotionLines Dataset): MELD is a conversational emotion recognition dataset that extends the original EmotionLines corpus by incorporating synchronized audio, visual, and textual modalities Poria et al. (2019). It contains approximately 1,433 multi-party dialogues

Table 8: Ablation comparing TTSD-FAR against plain TTSD (no stopping criterion) and TTSD-LS (loss-plateau stopping criterion) under progressively missing text conditions. We report F1 score on the BAH dataset as the proportion of unavailable textual input increases.

Method	Full	10%	20%	30%	40%	50%
No-Adaptation	0.6341	0.5905	0.5798	0.5710	0.5662	0.5614
TTSD	–	0.6065	0.6039	0.5887	0.5816	0.5728
TTSD-LS	–	0.5910	0.5820	0.5730	0.5645	0.5560
TTSD-FAR (Ours)	–	**0.6185**	**0.6120**	**0.5955**	**0.5910**	**0.5870**

and over 13,000 utterances extracted from the *Friends* TV series, each labeled with one of seven discrete emotions (Anger, Disgust, Sadness, Joy, Neutral, Surprise, Fear) and sentiment labels. The multimodal nature and conversational context make MELD a challenging benchmark for models that must integrate semantic, acoustic, and facial cues to infer emotion.

DFEW (Dynamic Facial Expression in the Wild): DFEW is a large-scale dynamic facial expression dataset collected from more than 1,500 movies and comprising over 16,000 short video clips depicting naturalistic expressions captured under unconstrained conditions Jiang et al. (2020). Unlike posed or laboratory-controlled datasets, DFEW includes significant variations in pose, illumination, occlusion, and actor demographics, with each clip annotated for one of seven basic emotions. This dataset serves as a benchmark for dynamic facial expression recognition (DFER) in the wild, emphasizing robustness to real-world noise and temporal dynamics.

BAH (Behavioral Ambivalence/Hesitancy Dataset): The BAH dataset is a recently introduced benchmark for recognizing ambivalence and hesitancy (A/H) in video recordings designed for behaviour change studies González-González et al. (2026). It contains 1,118 videos recorded from 224 participants across diverse demographics, captured while participants responded to stimuli intended to elicit ambivalence or hesitancy. Frame- and video-level annotations highlight segments containing A/H cues, and the dataset also provides aligned face crops, audio transcripts with timestamps, and participant metadata. BAH's focus on subtle and conflicting emotional states presents a distinct challenge compared to conventional discrete emotion recognition tasks, highlighting the need for models that can capture fine-grained behavioral cues across modalities and are able to perform well even under missing conditions. The dataset comes with a predefined train-test split. We present the results on the official test set. The performance measure reported is the F1 score.

Table 9: Summary of prompt variations for zero-shot inference with corresponding video-level F1 scores.

Prompt Type	**Prompt**	**F1 Score**
Simple	Classify the emotion in the video as either *Non-Ambivalent* or *Ambivalent*. Respond with only one word.	0.2827
Definition 1	Definition: Ambivalence is the state of having contradictory or conflicting feelings or attitudes towards something or someone simultaneously. Classify the emotion in the video as either *Non-Ambivalent* or *Ambivalent*. Respond with only one word.	0.3326
Definition 2	Definition: Ambivalence and hesitancy is understood as the simultaneous experience of desires for change and against change. Classify the emotion in the video as either *Non-Ambivalent* or *Ambivalent*. Respond with only one word.	0.3772
Transcript + Def 1	Video transcript: {transcript}. Definition: Ambivalence is the state of having contradictory or conflicting feelings or attitudes towards something or someone simultaneously. Classify the emotion in the video as either *Non-Ambivalent* or *Ambivalent*. Respond with only one word.	**0.6341**
Transcript + Def 2	Video transcript: {transcript}. Definition: Ambivalence and hesitancy are understood as the simultaneous experience of desires for change and against change. Classify the emotion in the video as either *Non-Ambivalent* or *Ambivalent*. Respond with only one word.	0.3945

IMPLEMENTATION DETAILS

MISSING MODALITY SIMULATION:

This section provides the details of the missing modality simulation for both vision and text missing conditions. The missing modality problem is a vast paradigm, and it is important to set the scope of the experimentation to effectively validate the method. The datasets used for the validation of the TTSD-FAR method all contain utterance-level annotations. For the text-missing scenario, the speaker utterance in the prompt is replaced by an empty string, and for the vision-missing scenario, zero-imputed frames tensor is fed into the model. Figure 9 shows the vision-missing and text-missing scenarios for the MELD dataset.



Figure 9: Illustration of the missing modality

MELD Dataset: We evaluate on the MELD test split using utterance-level metadata providing dialogue ID, speaker, transcription, and 7-class emotion labels (neutral, surprise, fear, sadness, joy, disgust, anger). Visual inputs are constructed from pre-extracted cropped and aligned facial frames, resized to 224×224 and uniformly sampled to $T = 8$, frames per utterance. Textual inputs are the raw utterance transcriptions from the metadata. Missing modality scenarios are simulated by replacing the utterance string with an empty string for text-missing conditions, controlled by a seeded random state (seed 42) for reproducibility. The model is prompted in instruction-format with the video and utterance as inputs, and asked to classify the emotion from the seven candidate labels. Student LoRA is adapted using the Adam optimizer with a learning rate 10^{-4} with a batch size of 1, where gradient accumulation over 8 steps is equivalent to performing a single effective update per sample, updating only the student LoRA parameters $r = 8$, $\alpha = 16$, and dropout 0.3. **DFEW Dataset:** Since Video-LLaVA-7B is used in all experiments, the data loading and prompting structure remains the same as MELD. Student LoRA is adapted using the Adam optimizer with a learning rate 10^{-4} with a batch size of 1, where gradient accumulation over 8 steps is equivalent to performing a single effective update per sample, updating only the student LoRA parameters $r = 8$, $\alpha = 16$, and dropout 0.3. All experiments use float16 precision. The results are reported on the Official Fold-1 test set.

BAH Dataset: Similar to DFEW and MELD datasets, we perform the experiments for the BAH dataset using the VideoLlava model. However, the task of Ambivalence/Hesitancy detection is more complex than basic emotion recognition. Following González-González et al. (2026), we use the prompting structure shown in Table 9 to get the best results. Table 9 shows high reliance on the textual input. Adding an accurate description of the task improves performance, and the best results are achieved with the definition and speaker utterance.

Student LoRA is adapted using the Adam optimizer with a learning rate 10^{-4} with a batch size of 1, where gradient accumulation over 8 steps is equivalent to performing a single effective update per sample, updating only the student LoRA parameters $r = 8$, $\alpha = 16$, and dropout 0.3. All experi-

ments use float16 precision. The results are reported on the official test set.

Compute details: For all experiments, the student is being aligned to the teacher's hidden representations, not its output distribution. We therefore fix generation to greedy decoding (do_sample=False) so that final predictions are deterministic given a fixed adapter state. Experiments were run on a machine with 4× ***NVIDIA A100-SXM4-40GB GPUs*** and CUDA version 12.8. The exact libraries and corresponding versions are included in the code supplement.

FAR PARAMETERS:

β_F, τ_{fc}, K, and τ_{fmi} together control the ACTIVE/ANCHORED state machine. β_F sets the EMA decay for the Fisher diagonal. τ_{fc} and K jointly gate consolidation: CI must stay below τ_{fc} for K consecutive adapted samples before the student is frozen. τ_{fmi} gates reactivation, triggering once FMI exceeds this threshold. M and ε play a minor role, affecting only reporting and numerical stability rather than the transition logic itself.

Table 10: FAR hyperparameters

Parameter	Value	Definition
β_F	0.99	EMA decay parameter for the online Fisher diagonal
τ_{fc}	0.02	Consolidation Index threshold
K	5	No. of consecutive adaptation steps required
τ_{fmi}	2.0	Fisher Mismatch Index threshold
ε	10^{-8}	Numerical stability constant