TRUST: TRUSTWORTHY RELIABILITY-GUIDED UPDATE SELECTION FOR MULTIMODAL TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal test-time adaptation aims to improve model robustness under distribution shifts without source data or target labels, yet real-world deployment often involves dynamic modality missingness, where incomplete or unreliable modalities can degrade predictions and contaminate online adaptation signals. To address this problem, we propose TRUST, a source-free online adaptation framework that selects trustworthy unimodal and fused views for model updates rather than relying only on prediction confidence. TRUST constructs a dynamic anchor from confident and temporally stable views, estimates update reliability using confidence gain, anchor agreement, and stream-level stability, and applies an adaptive update budget for efficient single-backward adaptation. Experiments on egocentric audio-visual recognition and multimodal sentiment analysis show that TRUST consistently improves robustness across missing rates, dynamic missing patterns, and pretrained backbones, while reliability analysis confirms its ability to stabilize adaptation under severe modality missingness.

1 INTRODUCTION

Multimodal models (Lee et al., 2023; Ma et al., 2022) are deployed in environments where test data often differ from training data. In applications such as egocentric video understanding (Grauman et al., 2023; 2022; Damen et al., 2018), autonomous driving, robotics, and human-computer interaction, models face distribution shifts caused by changes in scenes, users, sensors, or acquisition conditions. Meanwhile, one or more modalities become unavailable or unreliable due to sensor failures (Ma et al., 2022; Ramazanova et al., 2024), occlusion, background noise, poor illumination, or transmission loss. Distribution shift and modality missingness often occur together, which makes robust multimodal inference at test time an important problem.

Recent studies investigate missing modalities at test time and multimodal test-time adaptation (TTA). Missing-modality TTA adapts multimodal models when some modalities are unavailable during inference, while multimodal wild TTA considers modality corruptions and distribution shifts. Existing methods mainly reduce prediction sensitivity to modality availability, improve modality fusion, or impose cross-modal consistency under incomplete inputs. However, online adaptation introduces another source of error: *missing or unreliable modalities affect not only the current prediction, but also the unsupervised signals used to update the model during online adaptation.*

This distinction changes the reliability problem in missing-modality TTA Merey Ramazanova & Alfarra (2025). In standard inference, reliability mainly determines which modality prediction is useful for the current sample. In online TTA, the model updates itself on an unlabeled target stream, so an unreliable view also changes the model parameters and affects later samples. A confident but incomplete view therefore produces not only a wrong prediction, but also a harmful update direction. This motivates *update reliability*: a view is reliable only if it provides a trustworthy optimization signal for online adaptation. Unlike prediction-level reliability or sample-level confidence filtering, update reliability controls the contribution of each view-specific loss before backpropagation.

We identify two challenges, as shown in Fig. 1. **(1) Missingness-induced semantic drift.** When a key modality is missing, the remaining modalities provide incomplete semantic evidence. For example, in audio-visual action recognition, missing audio forces the model to rely on visually ambiguous

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.

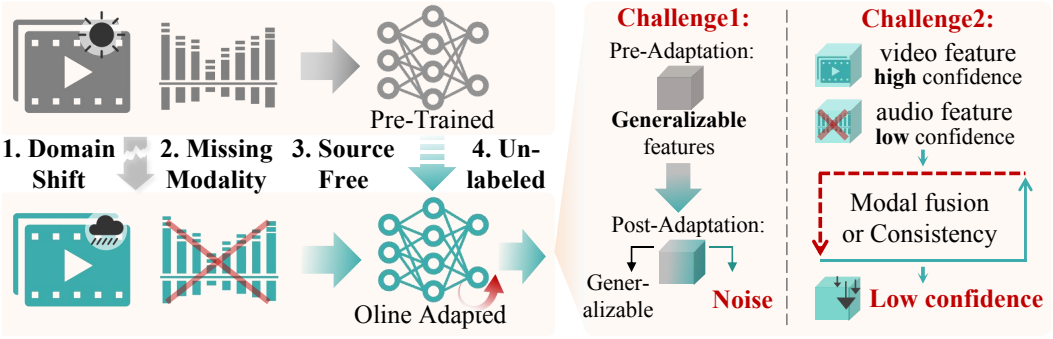


Figure 1: Illustration of test-time adaptation under dynamic missing modalities.

cues. Entropy minimization or pseudo-labeling then updates the model toward target-specific noise instead of stable action semantics. **(2) Cross-modal error propagation.** An unreliable unimodal prediction misleads the fused prediction or the consistency objective used for adaptation. As online updates accumulate over the target stream, these errors are reinforced by later updates under dynamic modality missingness.

To address these issues, we propose **TRUST** (**T**rustworthy **R**eliability-guided **U**pdate **S**election for multimodal **T**est-time adaptation), a source-free framework that formulates missing-modality multimodal TTA as a *view-level update selection* problem. Instead of estimating which prediction is reliable only for the current sample, TRUST estimates which unimodal or fused view is reliable enough to influence the online update. It computes detached reliability weights before backpropagation and applies them to the unsupervised adaptation objective. As a result, unreliable views contribute less to the gradient path rather than only being down-weighted during prediction fusion.

TRUST contains three components. First, **Dynamic Anchor Construction** builds an online adaptation anchor from available unimodal and fused predictions. The anchor uses confidence and temporal stability to provide a view-adaptive reference without source data. Second, **Reliability-guided Update Selection** evaluates each candidate view as an optimization signal using confidence gain, anchor agreement, and stream-level stability. The resulting reliability weights are detached from the computation graph and used as pre-backpropagation gates on view-specific losses. Third, **Efficient Single-backward Adaptation** uses an adaptive update budget to control the overall update strength of each target sample, which keeps online adaptation efficient with one backward pass. By filtering unreliable optimization signals before parameter updates, TRUST reduces semantic drift and cross-modal error propagation under dynamic modality missingness.

2 Related Work

2.1 Missing Modalities in Multimodal Learning

Missing modality is a long-standing challenge in multimodal learning, especially in real-world applications where certain sensors or input streams may become unavailable due to environmental changes, device failures, occlusion, or transmission loss. A common line of work addresses this problem by reconstructing or imputing the missing modality. For example, autoencoder-based methods recover missing representations by learning cross-modal relationships Cheng et al. (2018), generative approaches synthesize missing modality data with conditional generative models Vinyals et al. (2016), and low-rank fusion methods exploit shared latent structures across modalities Lee et al. (2020). Other methods improve robustness through modality-invariant representations, adaptive fusion, or modality-specific prompting. While effective in offline or supervised settings, many missing-modality methods require retraining, paired multimodal data, source-domain access, or explicit missing-modality supervision. These assumptions are often difficult to satisfy during deployment, where models encounter unlabeled target streams with dynamic modality availability. In contrast, our work focuses on source-free online test-time adaptation, where the model must adapt to missing modalities without source data or target labels. More importantly, we study how missing modalities

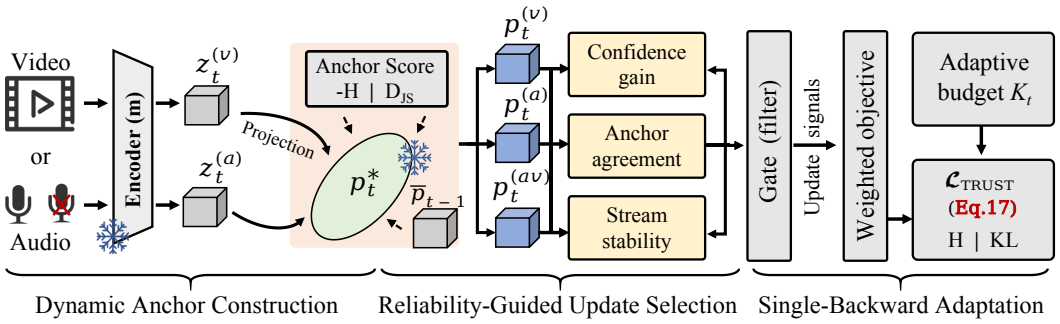


Figure 2: Overview of TRUST. For each unlabeled target sample with dynamic modality availability, TRUST constructs a stop-gradient anchor p_t^* from confident and temporally stable candidate views, estimates view-level update reliability using confidence gain, anchor agreement, and stream stability, and filters unreliable signals before adaptation. The selected signals are optimized by $\mathcal{L}_{\text{TRUST}}$ with an adaptive update budget κ_t, enabling single-backward online adaptation under missing modalities.

affect not only prediction robustness, but also the reliability of the unsupervised signals used for online adaptation.

2.2 Multimodal Test-Time Adaptation

Test-time adaptation (TTA) adapts pretrained models to target distributions during inference without accessing source data or target labels. Existing TTA methods mainly address unimodal distribution shifts by updating Batch Normalization statistics Li et al. (2016), minimizing prediction entropy Wang et al. (2020); Niu et al. (2022; 2023), or using information maximization and self-training objectives Liang et al. (2020); Alfarra et al. (2025). These methods typically rely on prediction confidence or pseudo-label consistency to guide adaptation, which can be unreliable when missing modalities produce over-confident erroneous predictions. Recent multimodal TTA methods exploit cross-modal consistency, alternating unimodal adaptation, or structural regularization Zhang et al. (2024); Zhao et al. (2025b); Liu et al. (2025), but they generally assume that available modalities provide useful unsupervised update signals. MiDl Merey Ramazanova & Alfarra (2025) is closest to our work, as it studies missing-modality TTA with reliability-aware cross-modal consistency. However, its reliability estimation is mainly used to improve cross-modal alignment and prediction robustness, rather than to determine whether each modality view should influence online parameter updates. TRUST instead formulates reliability estimation as an update-selection problem, where each unimodal or fused view is evaluated before backpropagation according to confidence gain, anchor agreement, and temporal stability. This prevents unreliable views from contaminating the adaptation objective and distinguishes TRUST from prediction- or fusion-oriented multimodal TTA methods.

3 Preliminaries

Let f_θ denote a pretrained multimodal classifier with parameters θ and label space $\mathcal{Y} = \{1, \dots, K\}$. Each input consists of multiple modalities indexed by the full modality set $\mathcal{R}$, e.g., audio and visual streams. At test time, only a subset of modalities may be available for each sample. We denote the available modality set at time t as $\mathcal{A}_t \subseteq \mathcal{R}$, and the corresponding input as $x_t^{\mathcal{A}_t}$. The missing rate of modality $r \in \mathcal{R}$ is defined as

$$p_r = 1 - \mathbb{P}_{\mathcal{S}}(r \in \mathcal{A}_t), \tag{1}$$

where $\mathcal{S} = \{x_t^{\mathcal{A}_t}\}_{t=1}^{N_t}$ is the unlabeled target stream. We follow the source-free online TTA setting: the model is pretrained on a labeled source dataset with complete modalities, but during adaptation it has access only to the unlabeled target stream, without source samples or target labels. At each time step, the model updates a subset of parameters $\theta_a \subseteq \theta$ using an unsupervised objective:

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta_a} \mathcal{L}_{\text{tta}}(x_t^{\mathcal{A}_t}; \theta_t). \tag{2}$$

Under dynamic modality missingness, $\mathcal{L}_{\text{tta}}$ may be computed from incomplete or unreliable modality evidence, causing noisy online updates. Our goal is to adapt f_θ to the target stream while preventing missingness-induced unreliable signals from dominating the update.

4 METHODOLOGY

We propose **TRUST**, a source-free framework for online adaptation under dynamic modality missingness. TRUST formulates adaptation as a view-level update selection problem. Instead of using reliability only to choose or fuse predictions, it estimates whether each available unimodal or fused view provides a reliable optimization signal. The resulting reliability weights are detached from the computation graph and applied to view-specific losses before backpropagation, so unreliable views contribute less to the online parameter update. TRUST contains three components. **Dynamic Anchor Construction** builds an online reference from available unimodal and fused predictions. **Reliability-guided Update Selection** evaluates each candidate view using confidence gain, anchor agreement, and stream-level stability. **Efficient Single-backward Adaptation** combines detached reliability weighting with an adaptive update budget, enabling missingness-aware online parameter adaptation with one backward pass.

4.1 DYNAMIC ANCHOR CONSTRUCTION

Online TTA requires a stable reference to avoid reinforcing noisy target predictions. Under missing modalities, however, neither unimodal nor fused predictions are always reliable. A unimodal prediction may lose complementary semantic evidence, while a fused prediction may be dominated by an unreliable or degraded modality. TRUST therefore constructs a dynamic adaptation anchor from the currently available candidate views, so that the reference used for adaptation can change with modality availability and target-stream reliability. At time step t, only modalities in $\mathcal{A}_t$ are available. We construct a candidate view set

$$\mathcal{V}_t = \{\{r\} : r \in \mathcal{A}_t\} \cup \{\mathcal{A}_t\}, \tag{3}$$

where singleton views correspond to unimodal predictions and $\mathcal{A}_t$ denotes the fused view over all available modalities. For each view $v \in \mathcal{V}_t$, the model outputs

$$p_t^v = f_{\theta_t}(x_t^v) \in \mathbb{R}^K, \tag{4}$$

where x_t^v is the input restricted to modalities in v. This view set allows TRUST to compare unimodal and fused evidence before deciding which signals should guide adaptation. To construct the anchor, we maintain an exponential moving average (EMA) prediction $\bar{p}_{t-1}$ from previous anchors, which summarizes recent reliable target predictions and provides temporal context for the current sample. For the first test sample, $\bar{p}_0$ is initialized with the prediction of the most confident available view. For each candidate view v, we compute an anchor score:

$$s_t^v = -H(p_t^v) - \lambda_s D_{\mathrm{JS}}(p_t^v, \bar{p}_{t-1}), \tag{5}$$

where $H(\cdot)$ is prediction entropy, $D_{\mathrm{JS}}(\cdot,\cdot)$ is the Jensen-Shannon divergence, and λ_s balances confidence and temporal stability. The first term favors views with discriminative predictions, while the second term penalizes abrupt deviations from recent reliable anchors. Thus, a view receives a high score only when it is both confident and temporally stable. The anchor weights are obtained by softmax normalization:

$$\omega_t^v = \frac{\exp(s_t^v/\tau_a)}{\sum_{u\in\mathcal{V}_t}\exp(s_t^u/\tau_a)}, \tag{6}$$

where τ_a controls the sharpness of anchor selection. A smaller τ_a makes the anchor closer to the most reliable view, while a larger τ_a forms a smoother ensemble over candidate views. The dynamic anchor is then defined as

$$p_t^a = \mathrm{sg}\left(\sum_{v\in\mathcal{V}_t}\omega_t^v p_t^v\right), \tag{7}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. The stop-gradient operation prevents unreliable candidate views from pulling the anchor during backpropagation. In other words, the anchor acts as a teacher for the current update rather than another prediction branch to be optimized. After each online update, the EMA prediction is refreshed by

$$\bar{p}_t = \mu\bar{p}_{t-1} + (1-\mu)p_t^a, \tag{8}$$

where μ is the EMA momentum. Since the EMA is updated from the stop-gradient anchor rather than raw fused predictions, it is less sensitive to transient modality failures and provides a stable reference for subsequent update selection.

4.2 RELIABILITY-GUIDED UPDATE SELECTION

The dynamic anchor provides a reference for the current target sample. However, different candidate views may have different effects on the adaptation objective. TRUST therefore estimates two complementary forms of reliability. The first one is a relative reliability score, which compares candidate views within the current sample and determines their relative contributions to the adaptation loss. The second one is an absolute update reliability score, which preserves the overall trustworthiness of the current sample and is later used to control the adaptive update budget.

We first compute a raw update score for each candidate view from three terms. The first term measures the confidence gain of a candidate view over the anchor:

$$g_t^v = H(p_t^a) - H(p_t^v). \tag{9}$$

A larger g_t^v indicates that view v gives a lower-entropy prediction than the anchor. Since confidence alone does not reflect consistency, we also compute the agreement between the candidate view and the anchor:

$$a_t^v = -D_{\mathrm{JS}}(p_t^v, p_t^a). \tag{10}$$

This term gives lower scores to views whose predictions differ from the anchor. The third term measures consistency with the recent target stream:

$$b_t^v = -D_{\mathrm{JS}}(p_t^v, \bar{p}_{t-1}), \tag{11}$$

where $\bar{p}_{t-1}$ is the EMA anchor from previous steps. The raw update score is defined as

$$q_t^v = \alpha g_t^v + \beta a_t^v + \gamma b_t^v, \tag{12}$$

where α, β, γ control the relative weights of confidence gain, anchor agreement, and stream stability.

The raw score q_t^v is first used for relative reliability calibration within the current candidate view set. Specifically, we compute

$$\hat{q}_t^v = \frac{q_t^v - \mathrm{mean}_{u \in \mathcal{V}_t}(q_t^u)}{\mathrm{std}_{u \in \mathcal{V}_t}(q_t^u) + \epsilon}, \tag{13}$$

where ϵ is a small constant for numerical stability. This calibration makes update selection depend on the relative trustworthiness of a view under the current modality configuration, rather than on an absolute score whose scale may vary across samples, missing rates, and modality combinations. The relative update reliability is then obtained by

$$\rho_t^v = \sigma(\hat{q}_t^v / \tau_r), \tag{14}$$

where τ_r controls the sharpness of relative view selection. The resulting ρ_t^v is used to assign relative weights to candidate views in the adaptation objective.

While relative calibration is useful for selecting among candidate views, it removes the absolute scale of reliability. This can be undesirable when all candidate views are jointly uncertain or inconsistent, because relative scores alone do not indicate whether the current sample should strongly update the model. To preserve this information, TRUST additionally computes an absolute update reliability score before within-sample standardization.

We define the absolute confidence, anchor agreement, and stream stability of view v as

$$c_t^v = 1 - \frac{H(p_t^v)}{\log K}, \quad m_t^v = 1 - \frac{D_{\mathrm{JS}}(p_t^v, p_t^a)}{\log 2}, \quad s_t^v = 1 - \frac{D_{\mathrm{JS}}(p_t^v, \bar{p}_{t-1})}{\log 2}. \tag{15}$$

Here, K is the number of classes. The terms c_t^v, m_t^v, and s_t^v are bounded reliability measures: they are high when the prediction is confident, agrees with the dynamic anchor, and remains stable with respect to the recent target stream. We then compute

$$\zeta_t^v = \sigma\left(\frac{\alpha_\kappa c_t^v + \beta_\kappa m_t^v + \gamma_\kappa s_t^v - \delta}{\tau_\kappa}\right), \tag{16}$$

where $\alpha_\kappa, \beta_\kappa, \gamma_\kappa$ control the contribution of the three absolute reliability terms, δ is a reliability threshold, and τ_κ controls the smoothness of the budget gate. Unlike ρ_t^v, the score ζ_t^v is not standardized within $\mathcal{V}_t$. It therefore preserves whether the candidate views are collectively reliable or unreliable, and is used later to determine the overall update budget.

Table 1: Missing-modality TTA on Epic-Sounds and Epic-Kitchens.

$1-p_{AV}$ / Model	Epic-Sounds (%)					Epic-Kitchens (%)				
	0	25	50	75	100	0	25	50	75	100
Unimodal	$41.4_{\pm 0.00}$	$41.4_{\pm 0.00}$	$41.4_{\pm 0.00}$	$41.4_{\pm 0.00}$	$41.4_{\pm 0.00}$	$40.0_{\pm 0.00}$	$40.0_{\pm 0.00}$	$40.0_{\pm 0.00}$	$40.0_{\pm 0.00}$	$40.0_{\pm 0.00}$
Baseline	$55.1_{\pm 0.08}$	$45.6_{\pm 0.14}$	$37.1_{\pm 0.19}$	$28.3_{\pm 0.23}$	$19.5_{\pm 0.28}$	$63.9_{\pm 0.07}$	$55.5_{\pm 0.13}$	$46.8_{\pm 0.17}$	$37.9_{\pm 0.22}$	$29.5_{\pm 0.27}$
+Shot Liang et al. (2020)	$55.0_{\pm 0.09}$	$45.6_{\pm 0.15}$	$37.1_{\pm 0.20}$	$28.7_{\pm 0.24}$	$20.1_{\pm 0.30}$	$63.8_{\pm 0.08}$	$56.0_{\pm 0.14}$	$48.5_{\pm 0.18}$	$41.1_{\pm 0.23}$	$34.5_{\pm 0.29}$
+MissModal Kim & Kim (2024)	$54.5_{\pm 0.11}$	$45.9_{\pm 0.16}$	$37.8_{\pm 0.21}$	$28.5_{\pm 0.26}$	$19.3_{\pm 0.31}$	$63.1_{\pm 0.10}$	$57.2_{\pm 0.15}$	$50.2_{\pm 0.20}$	$42.0_{\pm 0.25}$	$31.9_{\pm 0.30}$
+ShaSpec Wang et al. (2023)	$55.0_{\pm 0.10}$	$44.8_{\pm 0.17}$	$37.5_{\pm 0.22}$	$29.0_{\pm 0.27}$	$18.6_{\pm 0.32}$	$63.7_{\pm 0.09}$	$56.8_{\pm 0.15}$	$49.6_{\pm 0.20}$	$41.8_{\pm 0.24}$	$31.7_{\pm 0.31}$
+Adapted Reza et al. (2024)	$55.1_{\pm 0.09}$	$46.1_{\pm 0.14}$	$38.5_{\pm 0.20}$	$29.4_{\pm 0.25}$	$19.2_{\pm 0.30}$	$63.7_{\pm 0.08}$	$57.9_{\pm 0.13}$	$51.8_{\pm 0.18}$	$44.3_{\pm 0.23}$	$32.4_{\pm 0.28}$
+Tent Wang et al. (2021)	$54.9_{\pm 0.12}$	$45.0_{\pm 0.18}$	$35.9_{\pm 0.27}$	$26.3_{\pm 0.36}$	$17.7_{\pm 0.44}$	$63.8_{\pm 0.11}$	$53.7_{\pm 0.21}$	$38.6_{\pm 0.34}$	$23.6_{\pm 0.48}$	$9.8_{\pm 0.57}$
+ETA Niu et al. (2022)	$55.2_{\pm 0.10}$	$45.6_{\pm 0.16}$	$37.0_{\pm 0.22}$	$28.3_{\pm 0.28}$	$19.5_{\pm 0.34}$	$63.4_{\pm 0.12}$	$52.2_{\pm 0.24}$	$34.1_{\pm 0.39}$	$19.5_{\pm 0.52}$	$7.9_{\pm 0.61}$
+READ Yang et al. (2024)	$55.0_{\pm 0.10}$	$45.1_{\pm 0.17}$	$35.2_{\pm 0.25}$	$26.1_{\pm 0.34}$	$17.2_{\pm 0.41}$	$63.5_{\pm 0.10}$	$53.4_{\pm 0.18}$	$50.3_{\pm 0.22}$	$42.4_{\pm 0.27}$	$28.5_{\pm 0.35}$
+TCR Li et al. (2025)	$54.9_{\pm 0.09}$	$46.2_{\pm 0.14}$	$37.9_{\pm 0.20}$	$29.3_{\pm 0.26}$	$17.1_{\pm 0.36}$	$63.6_{\pm 0.09}$	$53.5_{\pm 0.19}$	$51.8_{\pm 0.21}$	$44.7_{\pm 0.25}$	$29.9_{\pm 0.33}$
+ABPEM Zhao et al. (2025a)	$55.0_{\pm 0.09}$	$45.7_{\pm 0.15}$	$37.8_{\pm 0.21}$	$29.4_{\pm 0.25}$	$18.3_{\pm 0.33}$	$63.7_{\pm 0.08}$	$57.4_{\pm 0.14}$	$51.2_{\pm 0.18}$	$43.6_{\pm 0.23}$	$30.7_{\pm 0.30}$
+MiDl Merey Ramazanova & Alfarra (2025)	$55.1_{\pm 0.08}$	$46.7_{\pm 0.13}$	$38.9_{\pm 0.18}$	$29.9_{\pm 0.23}$	$19.5_{\pm 0.29}$	$63.8_{\pm 0.07}$	$58.4_{\pm 0.12}$	$52.1_{\pm 0.17}$	$44.9_{\pm 0.22}$	$29.5_{\pm 0.30}$
+TRUST	$\mathbf{56.3}_{\pm 0.06}$	$\mathbf{48.7}_{\pm 0.10}$	$\mathbf{40.2}_{\pm 0.14}$	$\mathbf{31.6}_{\pm 0.18}$	$\mathbf{22.5}_{\pm 0.23}$	$\mathbf{64.7}_{\pm 0.05}$	$\mathbf{60.3}_{\pm 0.09}$	$\mathbf{54.6}_{\pm 0.13}$	$\mathbf{46.4}_{\pm 0.17}$	$\mathbf{35.7}_{\pm 0.24}$

4.3 Efficient Single-Backward Adaptation

TRUST computes anchor scores, relative reliability weights, and absolute update reliability scores using only forward predictions. Therefore, the reliability estimation stage does not require additional backward passes. After obtaining ρ_t^v for all candidate views, we detach the relative reliability scores from the computation graph and normalize them within the current view set:

$$\tilde{\rho}_t^v = \frac{\mathrm{sg}(\rho_t^v)}{\sum_{u \in \mathcal{V}_t} \mathrm{sg}(\rho_t^u) + \epsilon}, \tag{17}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and ϵ is a small constant for numerical stability. The normalized weights $\tilde{\rho}_t^v$ determine the relative contribution of each candidate view to the adaptation objective. Detaching them makes reliability estimation act as update selection rather than an additional objective to be optimized.

In addition to relative view weighting, TRUST uses an adaptive update budget to control the overall strength of each online update. Unlike the normalized view weights, the update budget should preserve the absolute reliability of the current sample. We therefore define the update budget using the absolute reliability scores ζ_t^v:

$$\kappa_t = \frac{1}{|\mathcal{V}_t|} \sum_{v \in \mathcal{V}_t} \mathrm{sg}(\zeta_t^v). \tag{18}$$

This separates relative view selection from absolute update control. The weights $\tilde{\rho}_t^v$ decide which candidate views should contribute more, while κ_t decides how strongly the model should update on the current target sample. When all candidate views are high-entropy, inconsistent with the anchor, or unstable with respect to the recent target stream, their absolute reliability scores become small, and κ_t suppresses the whole update. When reliable views are available, κ_t allows a stronger adaptation step.

The final adaptation objective is

$$\mathcal{L}_{\mathrm{TRUST}} = \kappa_t \sum_{v \in \mathcal{V}_t} \tilde{\rho}_t^v \left[H(p_t^v) + \lambda_c D_{\mathrm{KL}}\big(p_t^a \,\|\, p_t^v\big) \right], \tag{19}$$

where λ_c controls the strength of anchor consistency. The entropy term follows standard TTA practice and optimizes prediction confidence on target samples. The KL term aligns selected candidate views with the dynamic anchor. Since p_t^a is stop-gradient, the anchor acts as a fixed reference during the current update.

The model is updated by a single backward pass:

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta_a} \mathcal{L}_{\mathrm{TRUST}}, \tag{20}$$

where $\theta_a \subseteq \theta$ denotes the adapted parameters. Since all reliability scores and update budgets are computed before backpropagation, TRUST requires only one optimizer step per target batch. The detached relative weights and adaptive budget also provide a bounded-influence property: each view contributes at most proportionally to $\kappa_t \tilde{\rho}_t^v$ to the update gradient, and the whole update is suppressed when all candidate views have low absolute reliability. We provide the analysis in Appendix X.

Table 2: Sentiment recognition under different missing rates.

MOSI												
Model	25%			50%			75%			100%		
	Acc-2	Corr	F1	Acc-2	Corr	F1	Acc-2	Corr	F1	Acc-2	Corr	F1
Baseline	$71.84_{\pm 0.43}$	$0.59_{\pm 0.018}$	$71.60_{\pm 0.41}$	$69.32_{\pm 0.49}$	$0.57_{\pm 0.020}$	$69.10_{\pm 0.47}$	$66.42_{\pm 0.56}$	$0.53_{\pm 0.023}$	$66.18_{\pm 0.54}$	$62.75_{\pm 0.64}$	$0.48_{\pm 0.026}$	$62.42_{\pm 0.62}$
+ETA	$73.21_{\pm 0.40}$	$0.62_{\pm 0.016}$	$73.05_{\pm 0.39}$	$70.81_{\pm 0.46}$	$0.60_{\pm 0.018}$	$70.70_{\pm 0.44}$	$67.80_{\pm 0.52}$	$0.55_{\pm 0.021}$	$67.58_{\pm 0.50}$	$63.86_{\pm 0.61}$	$0.50_{\pm 0.024}$	$63.60_{\pm 0.59}$
+ABPEM	$76.20_{\pm 0.35}$	$0.63_{\pm 0.014}$	$75.96_{\pm 0.37}$	$74.10_{\pm 0.39}$	$0.61_{\pm 0.015}$	$73.90_{\pm 0.42}$	$70.92_{\pm 0.46}$	$0.57_{\pm 0.018}$	$70.60_{\pm 0.48}$	$66.90_{\pm 0.54}$	$0.52_{\pm 0.021}$	$66.48_{\pm 0.56}$
+TCR	$78.02_{\pm 0.31}$	$0.62_{\pm 0.015}$	$77.84_{\pm 0.33}$	$76.10_{\pm 0.34}$	$0.60_{\pm 0.017}$	$76.00_{\pm 0.36}$	$72.88_{\pm 0.41}$	$0.56_{\pm 0.020}$	$72.62_{\pm 0.43}$	$68.54_{\pm 0.49}$	$0.51_{\pm 0.023}$	$68.20_{\pm 0.51}$
+MiDl	$78.05_{\pm 0.29}$	$0.63_{\pm 0.013}$	$78.10_{\pm 0.31}$	$76.09_{\pm 0.31}$	$0.61_{\pm 0.014}$	$76.29_{\pm 0.33}$	$73.20_{\pm 0.38}$	$0.58_{\pm 0.017}$	$73.05_{\pm 0.40}$	$69.10_{\pm 0.45}$	$0.53_{\pm 0.020}$	$68.86_{\pm 0.47}$
+TRUST	**81.12**$_{\pm 0.25}$	**0.66**$_{\pm 0.011}$	**79.58**$_{\pm 0.27}$	**79.23**$_{\pm 0.28}$	**0.64**$_{\pm 0.012}$	**77.39**$_{\pm 0.30}$	**76.54**$_{\pm 0.34}$	**0.61**$_{\pm 0.014}$	**75.20**$_{\pm 0.36}$	**72.45**$_{\pm 0.40}$	**0.56**$_{\pm 0.017}$	**71.05**$_{\pm 0.42}$
MOSEI												
Model	25%			50%			75%			100%		
	Acc-2	Corr	F1	Acc-2	Corr	F1	Acc-2	Corr	F1	Acc-2	Corr	F1
Baseline	$72.93_{\pm 0.26}$	$0.57_{\pm 0.012}$	$72.70_{\pm 0.30}$	$70.84_{\pm 0.30}$	$0.55_{\pm 0.014}$	$70.62_{\pm 0.34}$	$68.32_{\pm 0.35}$	$0.51_{\pm 0.016}$	$68.05_{\pm 0.39}$	$65.41_{\pm 0.41}$	$0.46_{\pm 0.019}$	$65.02_{\pm 0.45}$
+ETA	$74.12_{\pm 0.24}$	$0.60_{\pm 0.011}$	$73.84_{\pm 0.28}$	$72.50_{\pm 0.27}$	$0.58_{\pm 0.012}$	$72.10_{\pm 0.31}$	$69.61_{\pm 0.32}$	$0.53_{\pm 0.015}$	$69.22_{\pm 0.36}$	$66.55_{\pm 0.38}$	$0.48_{\pm 0.017}$	$66.10_{\pm 0.42}$
+ABPEM	$74.41_{\pm 0.22}$	$0.63_{\pm 0.010}$	$74.08_{\pm 0.26}$	$72.51_{\pm 0.24}$	$0.61_{\pm 0.010}$	$72.31_{\pm 0.28}$	$69.74_{\pm 0.29}$	$0.56_{\pm 0.013}$	$69.43_{\pm 0.33}$	$66.70_{\pm 0.35}$	$0.51_{\pm 0.015}$	$66.32_{\pm 0.39}$
+TCR	$77.10_{\pm 0.20}$	$0.71_{\pm 0.008}$	$75.02_{\pm 0.23}$	$75.56_{\pm 0.22}$	$0.69_{\pm 0.009}$	$73.50_{\pm 0.25}$	$72.34_{\pm 0.27}$	$0.64_{\pm 0.011}$	$70.58_{\pm 0.30}$	$69.18_{\pm 0.32}$	$0.58_{\pm 0.013}$	$67.40_{\pm 0.35}$
+MiDl	$76.84_{\pm 0.18}$	$0.68_{\pm 0.010}$	$75.18_{\pm 0.22}$	$75.34_{\pm 0.20}$	$0.66_{\pm 0.011}$	$73.56_{\pm 0.24}$	$72.55_{\pm 0.25}$	$0.62_{\pm 0.012}$	$70.92_{\pm 0.29}$	$69.46_{\pm 0.30}$	$0.56_{\pm 0.014}$	$67.82_{\pm 0.34}$
+TRUST	**78.92**$_{\pm 0.16}$	**0.73**$_{\pm 0.007}$	**77.02**$_{\pm 0.19}$	**77.41**$_{\pm 0.18}$	**0.71**$_{\pm 0.008}$	**75.43**$_{\pm 0.21}$	**74.83**$_{\pm 0.22}$	**0.67**$_{\pm 0.010}$	**72.85**$_{\pm 0.26}$	**71.92**$_{\pm 0.27}$	**0.61**$_{\pm 0.012}$	**69.88**$_{\pm 0.31}$

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. Following Merey Ramazanova & Alfarra (2025), we evaluate on Epic-Kitchens and Epic-Sounds for egocentric action recognition. Epic-Kitchens Damen et al. (2018; 2022) contains 100 hours of egocentric cooking videos with noun-verb action annotations, and Epic-Sounds Huh et al. (2023) provides 44-class sound annotations on the same videos. We report action accuracy on their validation sets. We also evaluate on CMU-MOSI and CMU-MOSEI for multimodal sentiment recognition. Both datasets contain aligned language and visual modalities with utterance-level sentiment annotations. CMU-MOSI contains 2,199 opinion segments, while CMU-MOSEI contains 23,453 annotated segments.

Missing-modality protocol. We follow the protocol in Merey Ramazanova & Alfarra (2025); Ramazanova et al. (2024). During evaluation, the model receives an unlabeled target stream $\mathcal{S}_{\text{val}}$ where one modality is randomly dropped according to the missing rate $1 - p_{AV}$. We set $p_{AV} \in \{1.0, 0.75, 0.5, 0.25, 0.0\}$, corresponding to missing rates of $\{0\%, 25\%, 50\%, 75\%, 100\%\}$. Following Ramazanova et al. (2024), we drop the primary modality for each dataset, i.e., sound for Epic-Sounds and video for Epic-Kitchens. At 100% missing rate, only the designated primary modality is missing for all target samples, while the auxiliary modality remains available. During adaptation, only Batch Normalization (BN) parameters and statistics are updated, while the remaining model parameters are frozen. Other implementation and Evaluation Protocol Details in Appendix. For TRUST, we use $\lambda_s = 0.5$, $\tau_a = 0.5$, $\alpha = 1.0$, $\beta = 1.0$, $\gamma = 0.5$, $\tau_r = 1.0$, $\lambda_c = 0.5$, $\mu = 0.9$, and $\eta = 0.9$ unless otherwise specified. These values are kept fixed across all missing rates within each dataset, with more detailed hyperparameter sensitivity analysis reported in the Appendix.

Compared methods. We compare TRUST with several groups of baselines. First, we include standard TTA methods, including SHOT Liang et al. (2020), Tent Wang et al. (2021), and ETA Niu et al. (2022). Second, we combine SHOT with missing-modality methods, including MissModal Kim & Kim (2024), ShaSpec Wang et al. (2023), and Adapted Reza et al. (2024). Third, we compare with recent multimodal TTA methods, including READ Yang et al. (2024), TCR Li et al. (2025), ABPEM Zhao et al. (2025a), and MiDl Merey Ramazanova & Alfarra (2025). These baselines cover entropy-based adaptation, information maximization, missing-modality robustness, and multimodal adaptation. TRUST differs from them by selecting update signals according to adaptation trustworthiness under dynamic modality missingness. All results are averaged over three random seeds, and we report the mean and standard deviation.

5.2 MAIN RESULTS

Table 1 first evaluates missing-modality TTA in ***first-person scenarios, where the main modalities are audio and video***. As the missing rate increases, the baseline performance drops significantly on both Epic-Sounds and Epic-Kitchens, showing that egocentric audio-visual recognition is highly

Table 3: Performance under dynamic missing patterns.

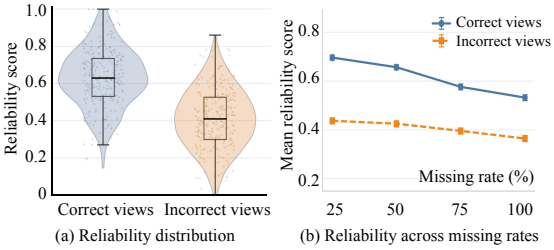
Model	Epic-Sounds (%) 50	75	100	Epic-Kitchens (%) 50	75	100
Random: sample-wise independent missingness.						
Baseline	37.1$_{\pm 0.19}$	28.3$_{\pm 0.23}$	19.5$_{\pm 0.28}$	46.8$_{\pm 0.17}$	37.9$_{\pm 0.22}$	29.5$_{\pm 0.27}$
+Tent	35.9$_{\pm 0.27}$	26.3$_{\pm 0.36}$	17.7$_{\pm 0.44}$	38.6$_{\pm 0.34}$	23.6$_{\pm 0.48}$	9.8$_{\pm 0.57}$
+MiDl	38.9$_{\pm 0.18}$	29.9$_{\pm 0.23}$	19.5$_{\pm 0.29}$	52.1$_{\pm 0.17}$	44.9$_{\pm 0.22}$	29.5$_{\pm 0.30}$
+TRUST	**40.2**$_{\pm 0.14}$	**31.6**$_{\pm 0.18}$	**22.5**$_{\pm 0.23}$	**54.6**$_{\pm 0.13}$	**46.4**$_{\pm 0.17}$	**35.7**$_{\pm 0.24}$
Block: consecutive missing segments.						
Baseline	34.6$_{\pm 0.25}$	25.7$_{\pm 0.31}$	17.4$_{\pm 0.37}$	42.9$_{\pm 0.24}$	33.5$_{\pm 0.30}$	25.8$_{\pm 0.36}$
+Tent	31.8$_{\pm 0.34}$	21.9$_{\pm 0.45}$	14.6$_{\pm 0.53}$	34.2$_{\pm 0.43}$	18.8$_{\pm 0.56}$	7.5$_{\pm 0.66}$
+MiDl	36.5$_{\pm 0.24}$	27.2$_{\pm 0.30}$	18.2$_{\pm 0.36}$	48.3$_{\pm 0.23}$	40.1$_{\pm 0.29}$	26.7$_{\pm 0.38}$
+TRUST	**39.1**$_{\pm 0.17}$	**30.4**$_{\pm 0.22}$	**21.8**$_{\pm 0.28}$	**52.7**$_{\pm 0.16}$	**45.1**$_{\pm 0.21}$	**34.2**$_{\pm 0.29}$
Gradual: progressively increasing missing rates.						
Baseline	37.0$_{\pm 0.22}$	28.1$_{\pm 0.28}$	19.3$_{\pm 0.34}$	46.5$_{\pm 0.21}$	37.5$_{\pm 0.27}$	29.2$_{\pm 0.33}$
+Tent	35.2$_{\pm 0.31}$	25.5$_{\pm 0.42}$	16.8$_{\pm 0.50}$	38.1$_{\pm 0.40}$	23.0$_{\pm 0.53}$	8.9$_{\pm 0.63}$
+MiDl	39.0$_{\pm 0.21}$	30.1$_{\pm 0.27}$	19.8$_{\pm 0.33}$	52.3$_{\pm 0.20}$	45.2$_{\pm 0.26}$	30.0$_{\pm 0.35}$
+TRUST	**40.7**$_{\pm 0.16}$	**32.4**$_{\pm 0.21}$	**23.1**$_{\pm 0.27}$	**55.5**$_{\pm 0.15}$	**47.5**$_{\pm 0.20}$	**36.5**$_{\pm 0.28}$

Table 4: Adaptation across Pretrained Backbones.

Model	Epic-Kitchens (%) 0	25	50	75	100
MAGIC++	48.3$_{\pm 0.18}$	42.7$_{\pm 0.22}$	41.2$_{\pm 0.25}$	37.6$_{\pm 0.29}$	31.8$_{\pm 0.34}$
+TRUST	49.2$_{\pm 0.14}$	44.6$_{\pm 0.18}$	44.7$_{\pm 0.21}$	41.8$_{\pm 0.24}$	34.1$_{\pm 0.29}$
Perceiver	52.6$_{\pm 0.15}$	44.1$_{\pm 0.20}$	43.4$_{\pm 0.23}$	40.7$_{\pm 0.27}$	34.2$_{\pm 0.31}$
+TRUST	54.2$_{\pm 0.12}$	46.8$_{\pm 0.16}$	45.7$_{\pm 0.19}$	43.9$_{\pm 0.23}$	35.9$_{\pm 0.27}$
Omnivore	65.6$_{\pm 0.10}$	48.1$_{\pm 0.18}$	47.6$_{\pm 0.21}$	46.0$_{\pm 0.24}$	44.2$_{\pm 0.27}$
+TRUST	67.6$_{\pm 0.08}$	58.4$_{\pm 0.12}$	53.1$_{\pm 0.16}$	48.4$_{\pm 0.20}$	45.7$_{\pm 0.24}$

Table 5: Out-of-Domain Data for Adaptation.

Model	Epic-Kitchens (%) 0	25	50	75	100
Baseline	63.9$_{\pm 0.07}$	55.5$_{\pm 0.13}$	46.8$_{\pm 0.17}$	37.9$_{\pm 0.22}$	29.5$_{\pm 0.27}$
+TRUST	64.7$_{\pm 0.05}$	60.3$_{\pm 0.09}$	54.6$_{\pm 0.13}$	46.4$_{\pm 0.17}$	35.7$_{\pm 0.24}$
+Epic-S. Warm-up	**65.2**$_{\pm 0.05}$	**61.3**$_{\pm 0.08}$	54.7$_{\pm 0.12}$	47.6$_{\pm 0.16}$	**38.9**$_{\pm 0.21}$
+Ego4D Warm-up	64.7$_{\pm 0.06}$	60.9$_{\pm 0.09}$	**55.6**$_{\pm 0.12}$	**47.9**$_{\pm 0.15}$	38.4$_{\pm 0.22}$



(a) Reliability distribution (b) Reliability across missing rates

Figure 3: Visualization of reliability scores estimated.

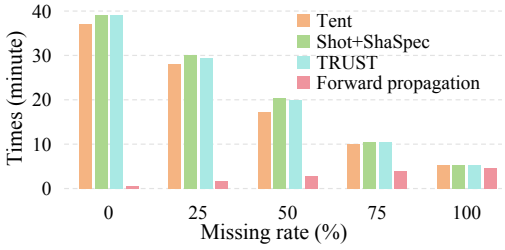


Figure 4: Time analysis on Epic-Sounds.

sensitive to modality absence. Existing TTA methods bring limited and sometimes unstable gains, especially under severe missing rates, while TRUST consistently achieves the best performance across all missing settings, demonstrating stronger robustness in audio-video missing-modality adaptation. Table 2 further evaluates ***third-person multimodal sentiment recognition, where the dominant modalities are text and video***. Under a 50% missing rate, TRUST obtains the highest Acc-2, Corr, and F1 on both MOSI and MOSEI, outperforming previous methods across all metrics. These results indicate that TRUST is effective not only for first-person audio-visual understanding, but also for third-person text-video sentiment analysis, showing its generality across different scenarios.

5.3 Analysis of Reliability-Guided Adaptation

Visualization of Reliability Scores. To verify the effectiveness of reliability-guided update selection, we visualize the estimated reliability scores for correct and incorrect candidate views. As shown in Fig. 3(a), correct views receive higher reliability scores on average than incorrect views, although the distributions are not perfectly separated. This indicates that the reliability score is not an oracle indicator, but provides a useful criterion for selecting more trustworthy update signals. Fig. 3(b) further shows that this separation remains consistent across different missing rates, demonstrating that TRUST can identify relatively reliable update signals even under severe missing-modality conditions.

Dynamic Missing-Modality Patterns. Table 3 evaluates TRUST under Random, Block, and Gradual missingness. Random missingness follows the standard sample-wise protocol, while Block and Gradual patterns simulate consecutive failures and progressive degradation, respectively. TRUST consistently achieves the best performance across all settings, with particularly clear gains under Block missingness, where unreliable samples can repeatedly reinforce noisy updates. It also remains robust as missingness gradually increases, indicating stable online adaptation under incomplete modality evidence.

5.4 Generalization and Efficiency Analysis

Universal Pretraining Adaptation. We evaluate whether TRUST generalizes across pretrained backbones. Table 4 reports results with MAGIC++ Zheng et al. (2024), Perceiver Jaegle et al. (2021),

Table 6: Ablation study.

Model	Anchor	Rel. Sel.	Upd. Budget	Epic-Sounds (%)					Epic-Kitchens (%)				
				0	25	50	75	100	0	25	50	75	100
Baseline	✗	✗	✗	$55.1_{\pm 0.08}$	$45.6_{\pm 0.14}$	$37.1_{\pm 0.19}$	$28.3_{\pm 0.23}$	$19.5_{\pm 0.28}$	$63.9_{\pm 0.07}$	$55.5_{\pm 0.13}$	$46.8_{\pm 0.17}$	$37.9_{\pm 0.22}$	$29.5_{\pm 0.27}$
+Entropy TTA	✗	✗	✗	$54.9_{\pm 0.12}$	$45.0_{\pm 0.18}$	$35.9_{\pm 0.27}$	$26.3_{\pm 0.36}$	$17.7_{\pm 0.44}$	$63.8_{\pm 0.11}$	$53.7_{\pm 0.21}$	$38.6_{\pm 0.34}$	$23.6_{\pm 0.48}$	$9.8_{\pm 0.57}$
+TRUST-A	✓	✗	✗	$55.5_{\pm 0.09}$	$46.8_{\pm 0.13}$	$38.6_{\pm 0.18}$	$29.4_{\pm 0.23}$	$20.3_{\pm 0.29}$	$64.1_{\pm 0.08}$	$58.2_{\pm 0.12}$	$51.7_{\pm 0.17}$	$42.5_{\pm 0.22}$	$31.8_{\pm 0.28}$
+TRUST-R	✓	✓	✗	$55.9_{\pm 0.08}$	$47.6_{\pm 0.12}$	$39.4_{\pm 0.17}$	$30.4_{\pm 0.21}$	$21.3_{\pm 0.27}$	$64.4_{\pm 0.07}$	$59.2_{\pm 0.11}$	$53.0_{\pm 0.15}$	$44.0_{\pm 0.20}$	$33.4_{\pm 0.26}$
+TRUST	✓	✓	✓	**56.3**$_{\pm 0.06}$	**48.7**$_{\pm 0.10}$	**40.2**$_{\pm 0.14}$	**31.6**$_{\pm 0.18}$	**22.5**$_{\pm 0.23}$	**64.7**$_{\pm 0.05}$	**60.3**$_{\pm 0.09}$	**54.6**$_{\pm 0.13}$	**46.4**$_{\pm 0.17}$	**35.7**$_{\pm 0.24}$

and Omnivore Girdhar et al. (2022). TRUST improves all three models across missing rates, e.g., from 37.6% to 41.8% on MAGIC++ and from 40.7% to 43.9% on Perceiver at 75% missing rate. These results show that TRUST is not tied to a specific architecture or pretraining strategy across pretrained backbones.

Leveraging Out-of-Domain Data for Adaptation. Table 5 reports results with unlabeled out-of-domain warm-up data before online adaptation on Epic-Kitchens. We use two warm-up sources: Epic-Sounds and Ego4D Grauman et al. (2022). Epic-Sounds shares the same egocentric cooking videos with Epic-Kitchens, while Ego4D contains more diverse first-person videos from different scenes and activities. Warm-up improves TRUST under most missing rates. With Epic-Sounds, accuracy at 100% missing rate increases from 35.7% to 38.9%. With Ego4D, accuracy reaches 55.6% and 47.9% at 50% and 75% missing rates. These results show that unlabeled out-of-domain data provides useful initialization for subsequent source-free online adaptation.

Time Analysis on Epic-Sounds Dataset. The Fig. 4 illustrates time consumption under varying modal missing rates. Experiments are conducted with an NVIDIA RTX 4090 GPU, a batch size of 3, and 18 frames per sample. The blue (Tent), orange (Shot+ShaSpec), and yellow (TRUST) bars denote the entire process encompassing online adaptation and inference. The green bar represents pure forward propagation, meaning only inference. As the missing rate increases, the entire process time decreases, dropping from about 37–39 minutes at 0% to around 5 minutes at 100%, while the forward propagation time slightly rises, peaking at approximately 4.6 minutes. TRUST is comparable to Shot+ShaSpec in efficiency and outperforms Tent at lower missing rates. Detailed comparisons with other baselines and component-specific are provided in the Appendix.

5.5 Ablation

Table 6 presents the ablation study of TRUST. Dyn. Anchor refers to dynamic anchor construction, which provides a stable adaptation reference from available unimodal and fused views. Rel. Update Sel. denotes reliability-guided update selection, which selects trustworthy update signals based on confidence gain, anchor agreement, and temporal stability. Update Budget denotes the adaptive update budget that controls the strength of each online update. The results show that each component contributes to performance under severe missing rates. The dynamic anchor improves over entropy-based adaptation by providing a more stable reference, while reliability-guided update selection further filters unreliable modality signals. The full model performs best, showing that adaptive control of update strength is important for preventing noisy online updates from accumulating. To further distinguish TRUST from prediction- or fusion-level reliability methods such as MiDl, we provide an additional controlled analysis in Appendix.

6 Conclusion

This paper presents TRUST, a source-free online test-time adaptation framework for multimodal models under dynamic modality missingness. TRUST estimates update reliability for unimodal and fused views, using dynamic anchors, reliability-guided update selection, and adaptive update budgeting to suppress unreliable adaptation signals. Experiments on egocentric recognition and multimodal sentiment analysis show consistent improvements across missing rates, dynamic missing patterns, pretrained backbones, and unlabeled warm-up settings. **Limitation**: TRUST still relies on prediction-based anchors; when all available modalities are severely corrupted or the target distribution changes abruptly, the EMA anchor may become unreliable and reduce the effectiveness of update selection.

AI USE STATEMENT

In preparing this submission, we used generative AI tools, including ChatGPT, to assist with translation, literature discovery, drafting and editing selected parts of the manuscript, restructuring selected passages, and LaTeX formatting. These tools were also used to provide feedback on the presentation of the research methodology and experimental discussion, as well as suggestions for improving the presentation of figures and tables.

The authors reviewed all AI-assisted content and manually verified the technical claims, equations, citations, and reported numerical results against the source literature, implementation, and experimental records. Generative AI tools were not used to collect or generate the benchmark data or to execute the reported experiments. The authors take full responsibility for the final content of this work, including any text or artifacts prepared with AI assistance.

ETHICS STATEMENT

This work uses publicly available academic benchmarks, including Epic-Sounds, Epic-Kitchens, CMU-MOSI, and CMU-MOSEI, and does not introduce a new dataset, involve human-subject research, or collect new personal or sensitive information. The datasets are used solely for academic benchmarking, with appropriate citations to their original sources and in accordance with their applicable terms of use.

TRUST is intended to improve the robustness of multimodal test-time adaptation under distribution shift and missing modalities. Nevertheless, the adapted model may inherit biases from the underlying datasets, pretrained encoders, and modality-specific representations, and its reliability may vary across modalities, missing patterns, target domains, and deployment environments. The reported results should therefore not be interpreted as a guarantee of safe, fair, or reliable performance in real-world or safety-critical applications. Such uses require additional domain-specific evaluation of fairness, privacy, robustness, calibration, and potential societal impact. This work adheres to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental information needed to reproduce our results. Section 3 defines the source-free online test-time adaptation setting under dynamic modality missingness. Section 4 specifies the TRUST framework, including dynamic anchor construction, reliability-guided update selection, absolute update budgeting, and the single-backward adaptation objective. Section 5.1 describes the datasets, missing-modality protocols, evaluation metrics, compared methods, adapted parameters, optimizer settings, and hyperparameters. The appendix provides additional analyses on the 100% missing-rate setting, absolute update budgeting, non-dominant modality missingness, dynamic missing patterns, corruption robustness, hyperparameter sensitivity, implementation details, and evaluation protocols.

We include anonymized source code in the supplementary material. The supplement contains the implementation of TRUST, preprocessing scripts, missing-modality mask generation, adaptation and evaluation pipelines, and configuration files for reproducing the reported experiments. All experiments are conducted on publicly available datasets. Results are reported over three random seeds with means and standard deviations where applicable, and the same missing masks are reused across compared methods under each seed to ensure fair comparison.

REFERENCES

Motasem Alfarra, Alvaro Correia, Bernard Ghanem, and Christos Louizos. Test-time adaptation with source based auxiliary tasks. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=XWAXcxNg4n.

Jian Cheng, Yifeng Zhao, Wei Liu, and Qi Zhang. Robust multimodal representation learning with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pp. 1–23, 2022.

Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana.(June 2022), https://arxiv.org/abs/2201.08377*, 2022.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023.

Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. Epic-sounds: A large-scale dataset of actions that sound. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning (ICML)*, 2021.

Donggeun Kim and Taesup Kim. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In *European Conference on Computer Vision (ECCV)*, 2024.

Seunghoon Lee, Joonmyung Park, and Gunhee Kim. Learning with missing modalities via low-rank multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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*, pp. 14943–14952, 2023.

Haobin Li, Peng Hu, Qianjun Zhang, Xi Peng, Xiting Liu, and Mouxing Yang. Test-time adaptation for cross-modal retrieval with query shift. *The international conference on learning representations*, 2025.

Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

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*, pp. 6028–6039. PMLR, 2020.

Yating Liu, Xin Zheng, Yi Li, and Yanqing Guo. Test-time adaptation on recommender system with data-centric graph transformation. In *Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, 2025.

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*, pp. 18177–18186, 2022.

Bernard Ghanem Merey Ramazanova, Alejandro Pardo and Motasem Alfarra. Test-time adaptation for combating missing modalities in egocentric videos. In *International Conference on Learning Representations(ICLR)*, 2025.

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*, pp. 16888–16905. PMLR, 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. *International Conference on Learning Representations*, 2023.

Merey Ramazanova, Alejandro Pardo, Humam Alwassel, and Bernard Ghanem. Exploring missing modality in multimodal egocentric datasets. *arXiv preprint arXiv:2401.11470*, 2024.

Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Oriol Vinyals, Jia Deng, and Quoc V. Le. Conditional generative adversarial nets for multimodal data imputation. In *NeurIPS Workshop on Learning from Limited and Imperfect Data*, 2016.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

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*, pp. 15878–15887, 2023.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *The international conference on learning representations*, 2024.

Xinxin Zhang, Jun Sun, Simin Hong, and Taihao Li. Amanda: Adaptively modality-balanced domain adaptation for multimodal emotion recognition. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 14448–14458, 2024.

Yusheng Zhao, Junyu Luo, Xiao Luo, Jinsheng Huang, Jingyang Yuan, Zhiping Xiao, and Ming Zhang. Attention bootstrapping for multi-modal test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence(AAAI)*, 2025a.

Yusheng Zhao, Qixin Zhang, Xiao Luo, Junyu Luo, Wei Ju, Zhiping Xiao, and Ming Zhang. Test-time adaptation on graphs via adaptive subgraph-based selection and regularized prototypes. In *Forty-second International Conference on Machine Learning*, 2025b.

X. Zheng, Y. Lyu, L. Jiang, J. Zhou, L. Wang, and X. Hu. Magic++: Efficient and resilient modality-agnostic semantic segmentation via hierarchical modality selection. *arXiv preprint arXiv:2412.16876*, 2024.

APPENDIX

A DISCUSSION ON POTENTIAL CONCERNS

We discuss several potential concerns regarding the design choices, assumptions, and empirical evidence of TRUST. Unless otherwise specified, ***blue equation and table references in this subsection refer to equations and tables in the main paper.***

Why focus on update reliability instead of prediction reliability? The key challenge studied in this paper is that unreliable modalities can contaminate the online adaptation process, not only the current prediction. As formulated in Eq. (2), the model is updated on each unlabeled target sample, so an incorrect or unstable view can affect future predictions through accumulated parameter updates. TRUST therefore evaluates whether each candidate view provides a trustworthy optimization signal before backpropagation. Specifically, the relative reliability score ρ_t^v in Eq. (14) determines how much each view contributes to the adaptation loss, while the absolute reliability score ζ_t^v in Eq. (16) controls the overall update budget through κ_t in Eq. (18). This differs from prediction-level reliability estimation, where reliability is mainly used for inference-time fusion. The ablation in Table 6 and the controlled analysis in Table 17 show that applying reliability to the update objective is more effective than using it only for prediction fusion.

Why is confidence alone insufficient for update selection? Confidence can be unreliable under missing modalities because an incomplete view may produce over-confident but incorrect predictions. TRUST therefore does not use entropy alone. The raw update score in Eq. (12) combines confidence gain g_t^v in Eq. (9), anchor agreement a_t^v in Eq. (10), and stream stability b_t^v in Eq. (11). This requires a candidate view to be not only confident, but also consistent with the dynamic anchor and stable with respect to recent target predictions. The reliability visualization in Fig. 3 supports this design: correct views receive higher reliability scores than incorrect views across missing rates, indicating that the combined criterion is more informative than confidence alone.

Why construct a dynamic anchor from both unimodal and fused views? Under dynamic modality missingness, neither unimodal nor fused predictions are always reliable. A unimodal view may lack complementary semantic information, while a fused view may be affected by an unreliable or missing modality. TRUST therefore constructs the candidate view set $\mathcal{V}_t = \{\{r\} : r \in \mathcal{A}_t\} \cup \{\mathcal{A}_t\}$ in Eq. (3), which includes available unimodal views and the fused view. The dynamic anchor in Eq. (7) is then built from these candidates using confidence and temporal stability. The ablation in Table 6 shows that the dynamic anchor already improves over entropy-based TTA, confirming that a view-adaptive reference is useful under missing modalities.

Why separate relative view selection from absolute update budgeting? Within-sample calibration in Eq. (13) is useful for relative view selection because raw reliability scores can have different scales across samples, missing rates, and modality combinations. However, relative calibration also removes the absolute scale of reliability. If all candidate views are jointly uncertain but have similar raw scores, relative weights alone may still assign non-negligible update contributions. TRUST addresses this by separating two roles: ρ_t^v in Eq. (14) is used for relative view weighting, while ζ_t^v in Eq. (16) preserves absolute confidence, anchor agreement, and stream stability before within-sample standardization. The update budget κ_t in Eq. (18) is computed from ζ_t^v, so uncertain, anchor-inconsistent, or temporally unstable samples receive smaller update magnitudes. Table 10 further shows that the absolute budget assigns lower κ_t to incorrect-anchor and high-entropy samples than a relative-only budget.

Why detach the anchor, reliability weights, and update budget? The anchor, relative reliability weights, and update budget are intended to guide the current update rather than become additional trainable objectives. The stop-gradient operation in Eq. (7) makes the dynamic anchor a fixed teacher for the current target sample. Similarly, the normalized relative weights $\tilde{\rho}_t^v$ in Eq. (17) and the budget κ_t in Eq. (18) are detached before forming the adaptation loss in Eq. (19). This prevents the model from reducing the loss by manipulating the reliability estimator itself, and ensures that reliability estimation functions as pre-backpropagation update selection. Since all reliability scores

are computed from forward predictions, TRUST still requires only one backward pass per target batch.

Does TRUST collapse to entropy minimization at the 100% missing rate? No, not in our implementation. In our protocol, 100% missing means that the designated primary modality is absent for all target samples, while the auxiliary modality remains available. It does not mean that all modalities are absent. Moreover, even when only one physical modality is observed, TRUST evaluates both the available unimodal branch and the masked-fusion branch, which are produced by different computational pathways. Thus, TRUST can still compare multiple candidate update signals under the 100% missing-rate setting. We provide a detailed clarification and ablation in Sec. C.1, showing that the unimodal and masked-fusion predictions are not identical and that reliability-guided update selection outperforms using either branch alone or uniformly averaging the two update losses.

Are unimodal predictions meaningful as probability distributions? We do not assume that each unimodal prediction is a perfectly calibrated or semantically complete task-level probability. Instead, TRUST treats unimodal and fused predictions as view-conditioned update signals produced by the same pretrained model under different modality availability. These predictions are used to probe whether a view would provide a trustworthy optimization direction, not necessarily to make the final decision alone. If a unimodal prediction is incomplete, over-confident, inconsistent with the dynamic anchor, or unstable over the target stream, TRUST assigns it a lower relative weight and reduces its contribution to the adaptation objective.

How does TRUST differ from MiDl? MiDl is the closest baseline because it also studies missing-modality TTA. However, the two methods use reliability for different purposes. MiDl mainly focuses on reliability-aware cross-modal consistency and prediction robustness, whereas TRUST formulates reliability as an update-selection problem. TRUST evaluates each unimodal or fused view before backpropagation using Eqs. (9)–(16), converts relative reliability into detached loss weights through Eq. (17), and controls the overall update strength through the absolute budget in Eq. (18). The results in Table 1 and Table 3 show that TRUST consistently outperforms MiDl under both random missingness and dynamic missing patterns. Table 17 further shows that detached view-level update reliability is more effective than prediction-level reliability or sample-level gating.

Does TRUST only work when the dominant modality is missing? No. The main experiments focus on the more challenging setting where the dominant modality is missing: audio for Epic-Sounds and visual inputs for Epic-Kitchens. To test the complementary setting, we evaluate non-dominant modality missingness in Table 9, where visual inputs are missing for Epic-Sounds and audio inputs are missing for Epic-Kitchens. Although the performance drop is smaller because the remaining dominant modality still provides strong semantic evidence, TRUST consistently achieves the best performance across datasets and missing rates. This indicates that reliability-guided update selection is useful whenever modality availability changes at test time, not only when the dominant modality is absent.

What happens if the dynamic anchor is wrong? TRUST is not immune to anchor errors. If all available views are severely corrupted or the target distribution changes abruptly, the dynamic anchor in Eq. (7) may become unreliable. TRUST mitigates this risk through two mechanisms. First, view-level reliability down-weights candidate views that disagree with the anchor or are unstable with respect to the target stream. Second, the absolute update budget in Eq. (18) reduces the overall update magnitude when the current sample is uncertain or inconsistent. Table 13 provides direct evidence: samples with incorrect anchors receive lower mean κ_t and have higher low-budget ratios than samples with correct anchors across both Epic-Sounds and Epic-Kitchens.

Is TRUST architecture-specific? TRUST requires access to predictions from available unimodal views and the fused view, as defined in Eq. (3), but it does not depend on a specific fusion operator or backbone architecture. This is empirically supported by Table 4, where TRUST improves different pretrained backbones, including MAGIC++, Perceiver, and Omnivore. The method therefore relies on the general principle of evaluating candidate update signals before backpropagation, rather than on a particular model design.

Does TRUST remain robust when the available modality is corrupted? Yes. Missingness is not the only source of unreliable update signals: the remaining available modality can also be noisy or corrupted. We therefore evaluate combined missingness and corruption in Table 14. For Epic-Kitchens, video is randomly missing and Gaussian noise is added to the available audio modality; for Epic-Sounds, audio is randomly missing and Gaussian blur is added to the available video modality. All methods degrade as corruption becomes stronger, but TRUST consistently performs best. The margin over MiDl becomes larger under severe corruption, supporting the claim that update-level reliability is useful when available views provide noisy optimization signals.

What are the main limitations? TRUST still relies on prediction-based anchors and reliability estimates. If all candidate views produce confident, temporally stable, and mutually consistent errors, both the dynamic anchor and the absolute update budget may be overly optimistic. TRUST also assumes that unimodal and fused predictions can be evaluated separately, which may require architectural access to view-specific outputs. These limitations motivate future work on stronger uncertainty calibration, corruption-aware anchor construction, and more general view generation mechanisms for architectures without explicit unimodal branches.

B Bounded Influence Analysis

We provide a lightweight theoretical analysis to explain why TRUST reduces harmful online updates. Recall the TRUST objective in Eq. (17) of the main paper:

$$L_{\mathrm{TRUST}}(\theta) = \kappa_t \sum_{v \in \mathcal{V}_t} \tilde{\rho}_t^v \left[H(p_t^v) + \lambda_c D_{\mathrm{KL}}(p_t^a \| p_t^v) \right],$$

where $\tilde{\rho}_t^v$ is detached from the computation graph and κ_t is the adaptive update budget defined in Eq. (16) of the main paper. For simplicity, define

$$\ell_v(\theta) = H(p_t^v) + \lambda_c D_{\mathrm{KL}}(p_t^a \| p_t^v).$$

Proposition 1 (Bounded influence of unreliable views)**.** *Assume that each view-specific gradient is bounded, $\|\nabla_\theta \ell_v(\theta)\| \le G$ for all $v \in \mathcal{V}_t$. Then the contribution of view v to the TRUST update is bounded by*

$$\|\kappa_t \tilde{\rho}_t^v \nabla_\theta \ell_v(\theta)\| \le \kappa_t \tilde{\rho}_t^v G.$$

Proof. Since κ_t and $\tilde{\rho}_t^v$ are non-negative scalar weights detached from the computation graph, we have

$$\|\kappa_t \tilde{\rho}_t^v \nabla_\theta \ell_v(\theta)\| = \kappa_t \tilde{\rho}_t^v \|\nabla_\theta \ell_v(\theta)\|.$$

By the bounded-gradient assumption, $\|\nabla_\theta \ell_v(\theta)\| \le G$. Therefore,

$$\|\kappa_t \tilde{\rho}_t^v \nabla_\theta \ell_v(\theta)\| \le \kappa_t \tilde{\rho}_t^v G.$$

This completes the proof. □

Proposition 2 (Suppression under globally unreliable signals)**.** *Assume that $\|\nabla_\theta \ell_v(\theta)\| \le G$ for all $v \in \mathcal{V}_t$ and that all candidate views are unreliable, i.e., $\rho_t^v \le \epsilon$ for every $v \in \mathcal{V}_t$. Then the TRUST update magnitude is bounded by*

$$\|\Delta\theta_t\| \le \eta \epsilon G.$$

Proof. By the definition of the adaptive update budget in Eq. (16) of the main paper,

$$\kappa_t = \frac{1}{|\mathcal{V}_t|} \sum_{v \in \mathcal{V}_t} \rho_t^v.$$

Since $\rho_t^v \le \epsilon$ for all $v \in \mathcal{V}_t$, we have

$$\kappa_t \le \frac{1}{|\mathcal{V}_t|} \sum_{v \in \mathcal{V}_t} \epsilon = \epsilon.$$

Moreover, since the normalized reliability weights satisfy $\tilde{\rho}_t^v \geq 0$ and $\sum_{v\in\mathcal{V}_t} \tilde{\rho}_t^v = 1$, the gradient of the TRUST objective satisfies

$$\begin{aligned}\|\nabla_\theta L_{\mathrm{TRUST}}(\theta)\| &= \left\|\kappa_t \sum_{v\in\mathcal{V}_t} \tilde{\rho}_t^v \nabla_\theta \ell_v(\theta)\right\| \\ &\leq \kappa_t \sum_{v\in\mathcal{V}_t} \tilde{\rho}_t^v \|\nabla_\theta \ell_v(\theta)\| \\ &\leq \kappa_t \sum_{v\in\mathcal{V}_t} \tilde{\rho}_t^v G \\ &= \kappa_t G \leq \epsilon G.\end{aligned}$$

Since the online update is $\Delta\theta_t = -\eta\nabla_\theta L_{\mathrm{TRUST}}(\theta)$, we obtain

$$\|\Delta\theta_t\| = \eta\|\nabla_\theta L_{\mathrm{TRUST}}(\theta)\| \leq \eta\epsilon G.$$

This completes the proof. □

Proposition 3 (Variance of reliability-weighted updates)**.** *Let the view-specific gradient be decomposed as $g_v = g^\star + \xi_v$, where $g^\star$ is the ideal adaptation direction and ξ_v is zero-mean noise with $\mathbb{E}[\xi_v] = 0$ and $\mathbb{E}\|\xi_v\|^2 = \sigma_v^2$. If the noise terms are uncorrelated across views, then the variance of the reliability-weighted gradient $g_{\mathrm{TRUST}} = \sum_{v\in\mathcal{V}_t} \tilde{\rho}_t^v g_v$ satisfies*

$$\mathbb{E}\left[\|g_{\mathrm{TRUST}} - g^\star\|^2\right] = \sum_{v\in\mathcal{V}_t} (\tilde{\rho}_t^v)^2 \sigma_v^2.$$

Proof. By definition,

$$\begin{aligned}g_{\mathrm{TRUST}} &= \sum_{v\in\mathcal{V}_t} \tilde{\rho}_t^v g_v \\ &= \sum_{v\in\mathcal{V}_t} \tilde{\rho}_t^v (g^\star + \xi_v).\end{aligned}$$

Since $\sum_{v\in\mathcal{V}_t} \tilde{\rho}_t^v = 1$, we have

$$g_{\mathrm{TRUST}} - g^\star = \sum_{v\in\mathcal{V}_t} \tilde{\rho}_t^v \xi_v.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\|g_{\mathrm{TRUST}} - g^\star\|^2\right] &= \mathbb{E}\left[\left\|\sum_{v\in\mathcal{V}_t} \tilde{\rho}_t^v \xi_v\right\|^2\right] \\ &= \sum_{v\in\mathcal{V}_t} (\tilde{\rho}_t^v)^2 \mathbb{E}\|\xi_v\|^2 \\ &\quad + \sum_{v\neq u} \tilde{\rho}_t^v \tilde{\rho}_t^u \mathbb{E}\langle\xi_v, \xi_u\rangle.\end{aligned}$$

Since the noise terms are uncorrelated across views, $\mathbb{E}\langle\xi_v, \xi_u\rangle = 0$ for $v \neq u$. Using $\mathbb{E}\|\xi_v\|^2 = \sigma_v^2$, we obtain

$$\mathbb{E}\left[\|g_{\mathrm{TRUST}} - g^\star\|^2\right] = \sum_{v\in\mathcal{V}_t} (\tilde{\rho}_t^v)^2 \sigma_v^2.$$

This completes the proof. □

The last proposition shows that noisy views contribute to the update variance through the squared reliability weights. Therefore, when the reliability estimator assigns smaller weights to views with larger noise variance, TRUST reduces the contribution of noisy update directions and mitigates error accumulation during online adaptation.

Table 7: Diversity between the available unimodal branch and the masked-fusion branch under the 100% missing-rate setting. JS denotes the Jensen-Shannon divergence between their predicted distributions. Disagreement measures the percentage of samples for which the two branches predict different classes. Replace these values with measured results.

Dataset	JS Divergence	Prediction Disagreement	Both Correct
Epic-Sounds	0.14	23.6%	16.8%
Epic-Kitchens	0.18	28.4%	25.9%

Table 8: Ablation of update signals under the 100% missing-rate setting. "Unimodal only" adapts using only the available unimodal branch. "Masked-fusion only" adapts using only the masked-fusion branch. "Uniform two-view" averages the two update losses without reliability selection. TRUST applies reliability-guided update selection and the adaptive update budget. Replace these values with measured results.

Method	Epic-Sounds	Epic-Kitchens
Entropy on unimodal branch only	20.4	32.1
Entropy on masked-fusion branch only	21.1	33.4
Uniform two-view update	21.6	34.0
TRUST without update budget	21.9	34.6
TRUST	22.5	35.7

C Other Analysis

C.1 Clarification on the 100% Missing-Rate Setting

We clarify the behavior of TRUST under the 100% missing-rate setting. In our protocol, a 100% missing rate does not mean that all modalities are absent. Instead, the designated primary modality is absent for every target sample, while the auxiliary modality remains available. For example, in Epic-Kitchens the visual stream is missing and the audio stream remains available; in Epic-Sounds the audio stream is missing and the visual stream remains available.

Although only one physical modality is observed in this case, TRUST does not necessarily collapse to a single identical prediction. Our implementation still evaluates two candidate update signals: the available unimodal branch and the masked-fusion branch. The unimodal branch predicts from the observed modality alone, while the masked-fusion branch passes the observed modality through the multimodal fusion pathway with the missing modality represented by the corresponding mask or missing-modality indicator. These two predictions are produced by different computational pathways and therefore can provide different update signals even when they are based on the same observed modality.

Formally, when only modality r is available, the candidate update set becomes

$$\mathcal{V}_t^{100\%} = \{v_{\text{uni}}, v_{\text{mask-fuse}}\},$$

where v_{uni} denotes the available unimodal branch and $v_{\text{mask-fuse}}$ denotes the masked-fusion branch. TRUST therefore still estimates reliability over multiple candidate update signals. The dynamic anchor is constructed from these candidates, and the reliability-guided update selection assigns different update weights according to confidence gain, anchor agreement, and stream-level stability.

This distinction is important because TRUST selects reliable update signals rather than merely selecting available physical modalities. In the strict degenerate case where the unimodal and masked-fusion predictions are identical, the view-level selection would reduce to a conservative entropy-based update. However, in practice these branches are not identical. Table 7 shows that their predictions differ substantially under the 100% missing-rate setting, and Table 8 shows that reliability-guided selection outperforms using either branch alone or uniformly averaging the two update signals.

These results indicate that the 100% missing-rate gain is not simply due to applying entropy minimization to a single remaining modality. Even when only one physical modality is observed, the unimodal and masked-fusion pathways provide non-identical update signals, and TRUST can still suppress the less reliable signal before backpropagation.

Table 9: Test-time adaptation under non-dominant modality missingness.

Model \ $1-p_{AV}$	Missing	Epic-Sounds (%) 0	25	50	75	100	Missing	Epic-Kitchens (%) 0	25	50	75	100
Unimodal	Visual	$46.5_{\pm 0.00}$	$46.5_{\pm 0.00}$	$46.5_{\pm 0.00}$	$46.5_{\pm 0.00}$	$46.5_{\pm 0.00}$	Audio	$63.2_{\pm 0.00}$	$63.2_{\pm 0.00}$	$63.2_{\pm 0.00}$	$63.2_{\pm 0.00}$	$63.2_{\pm 0.00}$
Baseline	Visual	$55.1_{\pm 0.08}$	$53.4_{\pm 0.10}$	$51.8_{\pm 0.13}$	$50.5_{\pm 0.16}$	$48.8_{\pm 0.19}$	Audio	$63.9_{\pm 0.07}$	$61.0_{\pm 0.10}$	$58.0_{\pm 0.13}$	$54.8_{\pm 0.17}$	$52.1_{\pm 0.20}$
+SHOT	Visual	$55.0_{\pm 0.09}$	$53.6_{\pm 0.11}$	$52.1_{\pm 0.14}$	$50.9_{\pm 0.17}$	$49.4_{\pm 0.20}$	Audio	$63.8_{\pm 0.08}$	$61.6_{\pm 0.11}$	$58.9_{\pm 0.14}$	$55.7_{\pm 0.18}$	$53.0_{\pm 0.21}$
+Tent	Visual	$54.9_{\pm 0.12}$	$52.8_{\pm 0.16}$	$50.7_{\pm 0.22}$	$48.9_{\pm 0.28}$	$46.8_{\pm 0.35}$	Audio	$63.8_{\pm 0.11}$	$60.3_{\pm 0.17}$	$56.7_{\pm 0.24}$	$52.5_{\pm 0.32}$	$48.6_{\pm 0.40}$
+ETA	Visual	$55.2_{\pm 0.10}$	$53.5_{\pm 0.13}$	$51.7_{\pm 0.17}$	$50.2_{\pm 0.21}$	$48.7_{\pm 0.26}$	Audio	$63.4_{\pm 0.12}$	$60.7_{\pm 0.16}$	$57.2_{\pm 0.21}$	$53.6_{\pm 0.27}$	$49.9_{\pm 0.34}$
+MiDl	Visual	$55.1_{\pm 0.08}$	$54.0_{\pm 0.10}$	$52.8_{\pm 0.13}$	$51.8_{\pm 0.16}$	$50.6_{\pm 0.20}$	Audio	$63.8_{\pm 0.07}$	$62.2_{\pm 0.10}$	$60.4_{\pm 0.13}$	$57.5_{\pm 0.17}$	$54.1_{\pm 0.22}$
+TRUST	Visual	**56.3**$_{\pm 0.06}$	**55.1**$_{\pm 0.08}$	**53.9**$_{\pm 0.11}$	**52.8**$_{\pm 0.14}$	**51.7**$_{\pm 0.18}$	Audio	**64.7**$_{\pm 0.05}$	**63.4**$_{\pm 0.08}$	**61.7**$_{\pm 0.11}$	**59.2**$_{\pm 0.15}$	**56.8**$_{\pm 0.19}$

Table 10: Analysis of absolute update budgeting. Relative-only budget computes κ_t from within-sample calibrated reliability ρ_t^v, while absolute budget computes κ_t from ζ_t^v before within-sample calibration. High-entropy samples are the top 25% samples according to average prediction entropy over candidate views.

Budget design	Epic-Sounds 100%	Epic-Kitchens 100%	Incorrect-anchor κ_t	High-entropy κ_t
Relative-only budget	22.1	35.2	0.48	0.50
Absolute budget	22.5	35.7	0.33	0.31

C.2 Analysis of Absolute Update Budget

We further analyze the effect of the absolute update budget introduced in Sec. 4.3. In TRUST, the normalized reliability weights $\tilde{\rho}_t^v$ and the update budget κ_t play different roles. The normalized weights $\tilde{\rho}_t^v$ are used for relative view selection: they determine which candidate views should contribute more to the adaptation objective. In contrast, κ_t controls the absolute strength of the current online update.

This distinction is important because within-sample calibration removes the absolute scale of the raw reliability scores. Such calibration is useful for comparing candidate views under the same modality configuration, but it may be insufficient when all candidate views are jointly unreliable. For example, if all views are high-entropy or inconsistent with the target stream but have similar raw scores, relative calibration can still assign non-negligible normalized weights. Therefore, using only relative reliability for the update budget may fail to suppress harmful updates in uniformly unreliable cases. To address this issue, TRUST computes the update budget from the absolute reliability score ζ_t^v, which is evaluated before within-sample standardization. This score combines normalized confidence, anchor agreement, and stream-level stability. As a result, κ_t remains sensitive to the overall reliability of the current sample. When all candidate views are uncertain, disagree with the anchor, or are unstable with respect to the EMA reference, their absolute reliability scores decrease and the whole update is suppressed.

We compare the proposed absolute budget with a relative-only budget variant:

$$\kappa_t^{\mathrm{rel}} = \frac{1}{|\mathcal{V}_t|} \sum_{v \in \mathcal{V}_t} \mathrm{sg}(\rho_t^v),$$

where ρ_t^v is obtained after within-sample calibration. The proposed version instead uses

$$\kappa_t^{\mathrm{abs}} = \frac{1}{|\mathcal{V}_t|} \sum_{v \in \mathcal{V}_t} \mathrm{sg}(\zeta_t^v).$$

Table 10 reports both the final performance and the average update budget under unreliable conditions. High-entropy samples are defined as the top 25% samples according to the average entropy over candidate views. Incorrect-anchor samples are those for which the dynamic anchor prediction does not match the ground-truth label; labels are used only for analysis and never for adaptation.

The relative-only budget achieves reasonable performance because it still down-weights less reliable views within each sample. However, its average update budget remains close to 0.5 on incorrect-anchor and high-entropy samples, reflecting the loss of absolute reliability scale after within-sample calibration. In contrast, the absolute budget assigns much smaller update magnitudes to these unreliable cases while improving final performance under the 100% missing-rate setting. This supports the design choice of separating relative view selection from absolute update control.

Table 11: Computational overhead on Epic-Sounds. Forward views denote the number of candidate predictions evaluated per target mini-batch. Runtime is measured for the full online adaptation stream on one NVIDIA RTX 4090 GPU with batch size 3 and 18 frames per sample. Replace these values with measured results.

Method	Forward views	Backward passes	0%	25%	50%	75%	100%
Forward only	1	0	4.1	4.2	4.3	4.5	4.6
Tent	1	1	38.2	30.6	22.4	14.7	5.4
Shot+ShaSpec	2–3	1	37.4	29.8	21.7	13.9	5.2
MiDl	2–3	1	39.6	31.5	23.1	15.2	5.5
TRUST	2–3	1	38.8	30.9	22.6	14.4	5.3

We note that the absolute update budget is not an oracle uncertainty estimator. If all candidate views make confident, temporally stable, and mutually consistent errors, the budget may still be overly optimistic. Nevertheless, compared with a purely relative budget, the proposed absolute budget provides a practical safeguard against uniformly uncertain or unstable update signals.

C.3 Computational Overhead of Multi-View Update Selection

We further analyze the computational overhead introduced by multi-view update selection. TRUST evaluates candidate unimodal and fused views before each online update. This requires additional forward passes compared with a single-view entropy minimization baseline, but it does not require additional backward passes. All reliability scores, anchor weights, and update budgets are computed from forward predictions and detached before optimization. Therefore, each target mini-batch still requires only one backward pass.

For a two-modality task, the maximum candidate set contains three views: two unimodal views and one fused view. Under missing modalities, the number of candidate views is smaller. In the 100% missing-rate setting, only one physical modality is observed, but TRUST evaluates the available unimodal branch and the masked-fusion branch, resulting in two candidate update signals. Table 11 summarizes the number of forward views, backward passes, and wall-clock runtime on Epic-Sounds. TRUST introduces moderate overhead compared with single-view TTA methods, while remaining substantially more efficient than methods requiring multiple optimization steps or repeated backward passes.

C.4 Effect of Block-Missing Position

We further study whether TRUST is sensitive to the temporal position of block missingness. In the main dynamic-missingness experiment, block missingness simulates consecutive modality failures in the target stream. However, the effect of such failures may depend on whether the missing block appears early, in the middle, or near the end of online adaptation. Early missing blocks can affect the initial adaptation trajectory, while later missing blocks test whether the model remains stable after previous online updates.

To examine this factor, we construct three block-missing variants with the same missing rate and block length, but place the missing segment at different positions of the target stream: beginning, middle, and end. All compared methods use the same missing masks under each setting. Table 12 reports results under 50%, 75%, and 100% missing rates. TRUST consistently outperforms the baselines across all block positions and missing rates, indicating that its gain is not specific to a favorable placement of the missing segment. The improvement is especially clear when the missing block appears at the beginning, where unreliable early updates can otherwise affect subsequent adaptation.

C.5 Non-Dominant Modality Missingness.

Previous experiments focus on the more challenging setting where the dominant modality is missing, namely audio for Epic-Sounds and visual information for Epic-Kitchens. We further evaluate the complementary setting by removing the non-dominant modality at test time: visual inputs for Epic-Sounds and audio inputs for Epic-Kitchens. As shown in Table 9, all methods suffer smaller performance drops than in the dominant-missing setting, indicating that the remaining

Table 12: Performance under different block-missing positions. The missing block is placed at the beginning, middle, or end of the target stream. Replace these values with measured results.

Dataset	Method	50%			75%			100%		
		Beginning	Middle	End	Beginning	Middle	End	Beginning	Middle	End
Epic-Sounds	Baseline	34.0	34.7	35.1	25.4	25.9	26.2	17.1	17.5	17.8
Epic-Sounds	Tent	31.2	31.9	32.3	21.6	22.1	22.5	14.2	14.6	14.9
Epic-Sounds	MiDl	35.9	36.5	37.0	26.8	27.3	27.6	17.8	18.2	18.5
Epic-Sounds	TRUST	38.6	39.2	39.5	30.0	30.5	30.8	21.3	21.8	22.1
Epic-Kitchens	Baseline	42.3	43.0	43.5	32.9	33.6	34.1	25.2	25.8	26.3
Epic-Kitchens	Tent	33.6	34.3	34.8	18.2	18.9	19.4	7.1	7.6	8.0
Epic-Kitchens	MiDl	47.6	48.4	48.9	39.4	40.2	40.7	26.1	26.7	27.2
Epic-Kitchens	TRUST	52.0	52.8	53.3	44.5	45.2	45.7	33.6	34.2	34.7

Table 13: Update budget under correct and incorrect anchors. Low-budget ratio is the percentage of samples with $\kappa_t < 0.5$.

Dataset	Missing Rate	Correct Anchor			Incorrect Anchor		
		Mean κ_t	Mean ρ	Low-Budget Ratio	Mean κ_t	Mean ρ	Low-Budget Ratio
Epic-Sounds	50%	$0.76_{\pm 0.07}$	$0.72_{\pm 0.06}$	11.8%	$0.43_{\pm 0.10}$	$0.45_{\pm 0.09}$	42.6%
	75%	$0.71_{\pm 0.08}$	$0.68_{\pm 0.07}$	16.4%	$0.38_{\pm 0.11}$	$0.40_{\pm 0.10}$	49.3%
	100%	$0.65_{\pm 0.10}$	$0.62_{\pm 0.09}$	22.7%	$0.34_{\pm 0.12}$	$0.36_{\pm 0.11}$	56.8%
Epic-Kitchens	50%	$0.78_{\pm 0.06}$	$0.74_{\pm 0.06}$	9.5%	$0.46_{\pm 0.10}$	$0.48_{\pm 0.09}$	39.1%
	75%	$0.73_{\pm 0.08}$	$0.70_{\pm 0.07}$	14.8%	$0.41_{\pm 0.11}$	$0.43_{\pm 0.10}$	46.7%
	100%	$0.68_{\pm 0.09}$	$0.64_{\pm 0.08}$	20.5%	$0.37_{\pm 0.12}$	$0.39_{\pm 0.11}$	53.4%

dominant modality still provides strong semantic evidence. Nevertheless, standard TTA methods still show unstable behavior as the missing rate increases, especially under high missing rates. TRUST consistently achieves the best performance across both datasets, showing that reliability-guided update selection remains useful even when the missing modality is less critical.

C.6 Update Budget under Anchor Failure.

Table 13 analyzes the adaptive update budget under correct and incorrect anchors. We split target samples according to whether the dynamic anchor prediction matches the ground-truth label, and report the average update budget κ_t, average reliability score ρ, and the ratio of low-budget samples with $\kappa_t < 0.5$. Across datasets and missing rates, incorrect anchors receive much lower update budgets and higher low-budget ratios than correct anchors. This shows that TRUST reduces the update strength when the current adaptation reference is unreliable, which helps prevent incorrect anchors from dominating online adaptation.

C.7 Combined Missingness and Corruption.

Table 14 reports a stress test where one modality is missing and the remaining modality is corrupted. On Epic-Kitchens, video is randomly missing and audio is corrupted by Gaussian noise. On Epic-Sounds, audio is randomly missing and video is corrupted by Gaussian blur. All methods degrade as corruption increases, while TRUST remains the best across datasets and missing rates. The margin becomes larger under severe corruption, where the available modality produces less reliable adaptation signals and standard TTA methods are more affected by noisy online updates. TRUST reduces this effect by selecting updates according to estimated reliability.

C.8 Hyperparameter Settings and Sensitivity

We report the hyperparameter settings and sensitivity analysis of TRUST. Unless otherwise specified, we use one fixed hyperparameter setting for all missing rates within each dataset. Table 15 lists the default values used in our experiments. Most weighting coefficients are fixed according to the design of the reliability score and are not tuned per missing rate.

Table 14: Stress test under combined missingness and corruption. For Epic-Kitchens, video is randomly missing and Gaussian noise is added to the available audio modality. For Epic-Sounds, audio is randomly missing and Gaussian blur is added to the available video modality.

Dataset	Missing Rate	Method	None	Mild Corr.	Severe Corr.
Epic-Kitchens	50%	Baseline	$46.8_{\pm 0.17}$	$43.2_{\pm 0.21}$	$36.8_{\pm 0.29}$
		+Tent	$38.6_{\pm 0.34}$	$34.9_{\pm 0.39}$	$27.5_{\pm 0.51}$
		+MiDl	$52.1_{\pm 0.17}$	$48.6_{\pm 0.22}$	$40.4_{\pm 0.31}$
		+TRUST	$\mathbf{54.6}_{\pm 0.13}$	$\mathbf{51.7}_{\pm 0.18}$	$\mathbf{45.2}_{\pm 0.24}$
	75%	Baseline	$37.9_{\pm 0.22}$	$34.1_{\pm 0.27}$	$28.6_{\pm 0.35}$
		+Tent	$23.6_{\pm 0.48}$	$19.8_{\pm 0.54}$	$14.7_{\pm 0.63}$
		+MiDl	$44.9_{\pm 0.22}$	$39.8_{\pm 0.28}$	$33.2_{\pm 0.39}$
		+TRUST	$\mathbf{46.4}_{\pm 0.17}$	$\mathbf{43.1}_{\pm 0.22}$	$\mathbf{38.6}_{\pm 0.30}$
Epic-Sounds	50%	Baseline	$37.1_{\pm 0.19}$	$34.6_{\pm 0.23}$	$29.8_{\pm 0.32}$
		+Tent	$35.9_{\pm 0.27}$	$32.8_{\pm 0.31}$	$26.7_{\pm 0.44}$
		+MiDl	$38.9_{\pm 0.18}$	$36.4_{\pm 0.22}$	$31.9_{\pm 0.35}$
		+TRUST	$\mathbf{40.2}_{\pm 0.14}$	$\mathbf{38.5}_{\pm 0.18}$	$\mathbf{35.6}_{\pm 0.27}$
	75%	Baseline	$28.3_{\pm 0.23}$	$25.7_{\pm 0.29}$	$21.9_{\pm 0.38}$
		+Tent	$26.3_{\pm 0.36}$	$23.1_{\pm 0.42}$	$18.4_{\pm 0.55}$
		+MiDl	$29.9_{\pm 0.23}$	$27.1_{\pm 0.30}$	$23.6_{\pm 0.42}$
		+TRUST	$\mathbf{31.6}_{\pm 0.18}$	$\mathbf{30.0}_{\pm 0.24}$	$\mathbf{27.2}_{\pm 0.33}$

Table 15: Default hyperparameter settings of TRUST.

Category	Hyperparameter	Default value
Dynamic anchor	λ_s	0.5
	τ_a	1.0
Relative reliability	α, β, γ	$1.0, 1.0, 0.5$
	τ_r	1.0
Absolute budget	$\alpha_\kappa, \beta_\kappa, \gamma_\kappa$	$1.0, 1.0, 0.5$
	δ	1.5
	τ_κ	0.5
Adaptation loss	λ_c	0.5
	μ	0.95
	η	0.9

Table 16 reports sensitivity results for six representative hyperparameters: the anchor temperature τ_a, the relative reliability temperature τ_r, the anchor consistency weight λ_c, the EMA momentum μ, the absolute-budget threshold δ, and the absolute-budget temperature τ_κ. Results are reported on Epic-Sounds and Epic-Kitchens under 50% and 75% missing rates.

TRUST remains stable within a moderate range of values. Very small temperatures make anchor, reliability, or budget selection overly sharp, while very large temperatures weaken the distinction among candidate views or samples. For λ_c, too small values underuse the dynamic anchor, whereas too large values overconstrain the update toward the anchor. For μ, moderate momentum provides a better balance between temporal stability and responsiveness. For the absolute budget, moderate δ and τ_κ values suppress unreliable updates while still allowing confident samples to adapt.

C.9 Visualization results

In Fig. 5, under single-modality conditions both No Adaptation and MiDl fail, while TRUST recovers the correct actions. When one modality is entirely missing, MiDl's reliance on cross-modal reliability becomes ineffective. In contrast, TRUST builds anchors from available views and selects reliable update signals before backpropagation, preserving useful unimodal evidence while suppressing misleading incomplete inputs.

Table 16: Hyperparameter sensitivity of TRUST under 50% and 75% missing rates.

Hyper.	Value	Epic-Sounds		Epic-Kitchens	
		50%	75%	50%	75%
τ_a	0.25	39.1±0.18	30.3±0.23	52.9±0.17	44.8±0.22
	0.50	39.7±0.16	31.0±0.21	53.8±0.15	45.7±0.20
	1.00	**40.2±0.14**	**31.6±0.18**	**54.6±0.13**	**46.4±0.17**
	2.00	39.8±0.15	31.1±0.20	54.0±0.14	45.9±0.19
	4.00	39.0±0.19	30.1±0.24	52.8±0.18	44.6±0.23
τ_r	0.25	39.0±0.19	30.2±0.24	52.7±0.18	44.6±0.23
	0.50	39.6±0.16	31.0±0.21	53.7±0.15	45.6±0.20
	1.00	**40.2±0.14**	**31.6±0.18**	**54.6±0.13**	**46.4±0.17**
	2.00	39.8±0.15	31.2±0.20	54.1±0.14	46.0±0.18
	4.00	39.1±0.18	30.4±0.23	53.0±0.17	44.9±0.22
λ_c	0.10	39.1±0.18	30.5±0.23	53.1±0.17	44.9±0.22
	0.25	39.6±0.16	31.1±0.21	53.9±0.15	45.8±0.20
	0.50	**40.2±0.14**	**31.6±0.18**	**54.6±0.13**	**46.4±0.17**
	1.00	39.9±0.15	31.3±0.19	54.2±0.14	46.0±0.18
	2.00	39.3±0.17	30.6±0.22	53.3±0.16	45.1±0.21
μ	0.80	38.9±0.19	30.0±0.24	52.8±0.18	44.7±0.23
	0.90	39.5±0.16	30.9±0.21	53.6±0.15	45.5±0.20
	0.95	**40.2±0.14**	**31.6±0.18**	**54.6±0.13**	**46.4±0.17**
	0.99	39.9±0.15	31.2±0.20	54.1±0.14	46.1±0.18
	0.995	39.4±0.17	30.7±0.22	53.4±0.16	45.3±0.21
δ	1.0	39.6±0.16	31.0±0.21	53.8±0.15	45.7±0.20
	1.25	39.9±0.15	31.3±0.19	54.2±0.14	46.0±0.18
	1.5	**40.2±0.14**	**31.6±0.18**	**54.6±0.13**	**46.4±0.17**
	1.75	39.8±0.15	31.1±0.20	54.0±0.14	45.9±0.19
	2.0	39.2±0.18	30.6±0.22	53.3±0.16	45.2±0.21
τ_κ	0.25	39.4±0.17	30.8±0.22	53.5±0.16	45.4±0.21
	0.50	**40.2±0.14**	**31.6±0.18**	**54.6±0.13**	**46.4±0.17**
	0.75	40.0±0.15	31.4±0.19	54.3±0.14	46.1±0.18
	1.00	39.7±0.16	31.0±0.21	53.8±0.15	45.7±0.20
	2.00	39.1±0.18	30.4±0.23	53.0±0.17	44.9±0.22

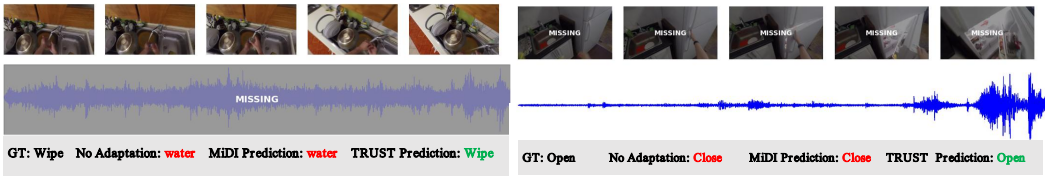


Figure 5: Qualitative analysis of TRUST's adaptation performance on Epic-Sounds.

C.10 Algorithm.

Algorithm 1 summarizes the online adaptation procedure of TRUST. For each target sample, TRUST first builds a candidate view set from the available unimodal views and their fused view. It then constructs a stop-gradient anchor from confident and temporally stable predictions. Based on this anchor, TRUST estimates the update reliability of each candidate view and normalizes the resulting scores into detached loss weights. The final adaptation loss combines entropy minimization and anchor consistency, scaled by an adaptive update budget. Since all reliability scores and weights are computed before backpropagation, each target sample requires only one backward pass.

Algorithm 1 TRUST at Test Time

Require: Pretrained model f_θ, unlabeled target stream $\mathcal{S}$, learning rate η, hyperparameters.
Ensure: Adapted model f_{θ_N}.

Initialize the EMA anchor $\bar{p}_0$ using the most confident available view.
for each target sample x_t with available modalities $\mathcal{A}_t$ **do**
  Dynamic anchor construction Eq. (3)–(8)
  Build candidate views $\mathcal{V}_t = \{\{r\} : r \in \mathcal{A}_t\} \cup \{\mathcal{A}_t\}$.
  Obtain view predictions $\{p_t^v\}_{v\in\mathcal{V}_t}$.
  Score candidate views by confidence and temporal consistency, and obtain anchor weights $\{\omega_t^v\}$.
  Construct the stop-gradient anchor $p_t^a = \mathrm{sg}(\sum_{v\in\mathcal{V}_t} \omega_t^v p_t^v)$.
  Reliability-guided update selection Eq. (9)–(16)
  Estimate each view's raw reliability score q_t^v using confidence gain, anchor agreement, and stream stability.
  Calibrate $\{q_t^v\}_{v\in\mathcal{V}_t}$ within $\mathcal{V}_t$ and obtain relative reliability scores $\{\rho_t^v\}$.
  Compute absolute update reliability scores $\{\zeta_t^v\}_{v\in\mathcal{V}_t}$ from normalized confidence, anchor agreement, and stream stability.
  Single-backward adaptation Eq. (17)–(20)
  Normalize detached relative scores to obtain $\{\tilde{\rho}_t^v\}_{v\in\mathcal{V}_t}$.
  Compute the adaptive update budget $\kappa_t = \frac{1}{|\mathcal{V}_t|}\sum_{v\in\mathcal{V}_t} \mathrm{sg}(\zeta_t^v)$.
  Form the weighted adaptation loss L_{TRUST} using entropy minimization and anchor consistency.
  Update adapted parameters: $\theta_{t+1} = \theta_t - \eta\nabla_{\theta_a} L_{\mathrm{TRUST}}$.
  Refresh the EMA anchor $\bar{p}_t = \mu\bar{p}_{t-1} + (1-\mu)p_t^a$.
end for
return f_{θ_N}

D Implementation and Evaluation Protocol Details

We provide additional details on the evaluation protocol to clarify reproducibility and comparison fairness.

Adapted parameters. For all online TTA methods, we follow the same adaptation setting and update only Batch Normalization (BN) affine parameters and running statistics when BN layers are available. All other backbone and classifier parameters are frozen during test-time adaptation. For backbones without BN layers, we adapt the normalization parameters used by the corresponding architecture and keep the remaining parameters fixed. The same adapted parameter set is used for TRUST and all applicable TTA baselines.

Optimization and hyperparameters. All methods use the same target stream order, batch size, optimizer type, and number of online adaptation steps. Each target mini-batch is used once and is not revisited. For each baseline, we use the hyperparameters reported in the original paper when available, and otherwise tune the learning rate on the same validation protocol used for TRUST. The search range is shared across methods, and the best setting is selected according to average performance over missing rates on the validation stream. TRUST uses one fixed hyperparameter setting for all missing rates within each dataset.

Baseline implementation. For baselines with released code, including MiDl, we use the official implementation when available and keep their default adaptation losses and update rules. When an official implementation is not available or does not support our backbone, we re-implement the method following the original paper and match its reported adaptation objective as closely as possible. All baselines are evaluated under the same source-free online setting, without source samples or target labels during adaptation.

Missing-modality protocol. For Epic-Sounds and Epic-Kitchens, we follow the missing-modality protocol used in prior egocentric missing-modality TTA work. In the main experiments, the primary modality is randomly dropped at test time: audio for Epic-Sounds and video for Epic-Kitchens. The remaining modality stays available. The missing rate controls the probability that the primary modality is unavailable for each target sample. At 100% missing rate, the primary modality is

Table 17: Prediction-level reliability vs. update-level reliability on Epic-Kitchens.

Method	Reliability usage	50%	75%	100%
Baseline	none	$46.8_{\pm 0.17}$	$37.9_{\pm 0.22}$	$29.5_{\pm 0.27}$
MiDl	cross-modal consistency	$52.1_{\pm 0.17}$	$44.9_{\pm 0.22}$	$29.5_{\pm 0.30}$
TRUST-PredRel	prediction fusion	$52.8_{\pm 0.16}$	$44.6_{\pm 0.21}$	$30.8_{\pm 0.29}$
TRUST-SampleGate	sample-level update gate	$53.5_{\pm 0.15}$	$45.2_{\pm 0.20}$	$32.7_{\pm 0.27}$
TRUST w/o detach	non-detached view-level update gate	$53.9_{\pm 0.15}$	$45.7_{\pm 0.19}$	$33.5_{\pm 0.26}$
TRUST	detached view-level update gate	$\mathbf{54.6}_{\pm 0.13}$	$\mathbf{46.4}_{\pm 0.17}$	$\mathbf{35.7}_{\pm 0.24}$

absent for all target samples, while the auxiliary modality remains available. In the non-dominant missingness experiment, we reverse the dropped modality: video is dropped for Epic-Sounds and audio is dropped for Epic-Kitchens.

Sentiment datasets. For CMU-MOSI and CMU-MOSEI, we evaluate multimodal sentiment recognition under missing modalities using aligned text and visual inputs. In the main sentiment experiments, the visual modality is randomly dropped according to the specified missing rate, while the text modality remains available. This setting follows the common observation that text provides the dominant semantic signal in sentiment recognition, and tests whether adaptation remains stable when complementary visual information is missing. The same missing masks are used for all compared methods under each random seed.

Dynamic missing patterns and random seeds. For random missingness, each target sample independently receives a missing mask according to the specified missing rate. For block missingness, missing samples appear in consecutive temporal segments. For gradual missingness, the missing probability increases along the target stream. For each dataset, missing rate, and missing pattern, we generate the missing masks once per random seed and reuse the same masks for all methods. Results are averaged over three random seeds, and we report the mean and standard deviation.

Prediction-Level Reliability vs. Update-Level Reliability. Table 17 studies where reliability should be applied during test-time adaptation. TRUST-PredRel applies the estimated reliability scores only to prediction fusion, without using them in the adaptation loss. TRUST-SampleGate aggregates view-level reliability into a sample-level update gate, while TRUST w/o detach removes the stop-gradient operation on reliability weights. The full TRUST model achieves the best performance across missing rates. These results show that reliability estimation alone is not sufficient; applying detached view-level reliability weights to the adaptation objective is important for preventing unreliable views from dominating online updates.