REVISITING GEOMETRIC SYMMETRY NON-CONSERVATION FOR MULTIMODAL TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Without access to source data, multimodal test-time adaptation (MTTA) has emerged as a promising paradigm for adapting multimodal models to target domains under complex distribution shift. Despite the documented success of MTTA, an overlooked yet persistent issue remains, which we term *geometric symmetry non-conservation*. This phenomenon refers to the breakdown of pretraining-established geometric symmetry between representations of different modalities under complex distribution shift. We uncover the cause of geometric symmetry non-conservation through a graphical explanation and a theoretical derivation. Furthermore, we propose ***ManiGest***, short for ***Mani****fold-constrained* ***G****eometry* ***Rest****oration*, to alleviate the impact of complex distribution shift on the geometric symmetry between representations of different modalities. ManiGest restores the geometric symmetry under the constraint of probabilistic Birkhoff polytope manifold, and leverages the Riemannian distance on the manifold of symmetric positive definite matrices to preserve intra-modal geometry. Extensive results on multiple benchmarks demonstrate that ManiGest serves as an effective plug-and-play module for improving robustness against complex distribution shift.

1 INTRODUCTION

Multimodal models Gong et al. (2023); Radford et al. (2021); Song et al. (2026), following the pretraining then fine-tuning Mu et al. (2025) paradigm, have shown substantial potential across diverse applications. However, their effectiveness on downstream tasks typically hinges on the assumption that the source and target domains are identically distributed Chen et al. (2023), which is frequently violated due to distribution shifts in practice. While numerous unimodal test-time adaptation (UTTA) methods Wang et al. (2021); Shin et al. (2022); Niu et al. (2022; 2023) improve robustness against distribution shifts, their effectiveness often deteriorates in multimodal scenarios because they treat different modalities equally under complex distribution shifts Yang et al. (2024). Concretely, in practical multimodal scenarios, distribution shift does not necessarily affect all modalities, but instead occurs only in certain modalities, e.g., in nighttime autonomous driving, dark light only affects camera signals, whereas LiDAR data remains immune. We refer to this complex distribution shift in multimodal scenarios as *modal-specific distribution shift*[1]. This exacerbates the information discrepancies between modalities and induces reliability bias across modalities Yang et al. (2024), and various multimodal test-time adaptation (MTTA) methods Yang et al. (2024); Chen et al. (2025b); Zhao et al. (2025) are designed to remedy this issue.

Most of these MTTA methods focus on optimizing the attention-based fusion Vaswani et al. (2017) module through variants of entropy minimization Wang et al. (2021) or meticulously designed regularization terms. Despite their remarkable performance, we identify an overlooked yet persistent representation-level issue impeding the further development of MTTA, which we term *geometric symmetry non-conservation*. It describes the failure to preserve the geometric symmetry established

[1] A prior study Chen et al. (2025b) also refers to this complex distribution shift as *unimodal distribution shift*, but we remove the term "unimodal distribution shift" from our consideration for the possible ambiguity with distribution shifts in purely unimodal scenarios. The complex distribution shift mentioned in **Abstract** specifically refers to the modal-specific distribution shift.

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.

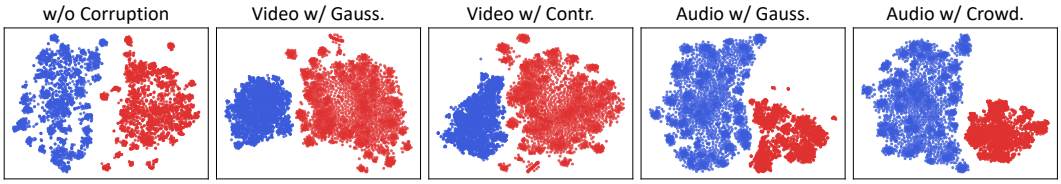

Figure 1: Illustration of geometric symmetry non-conservation. We utilize the video and audio encoders of CAV-MAE Gong et al. (2023) to extract representations of samples in the Kinetics50 Kay et al. (2017) dataset. The blue and red points denote the representations from the video and audio modalities, respectively. Following the widely adopted MTTA baselines Yang et al. (2024); Zhao et al. (2025), the modal-specific distribution shift is artificially constructed by introducing corruption into either audio or video modality Hendrycks & Dietterich (2019). *Gauss.*, *Contr.*, and *Crowd.* are three types of corruption detailed in Yang et al. (2024). The extended experimental results on a total of **21** corruptions for both video and audio, presented in **Appendix** A.3, confirm that geometric symmetry non-conservation is a consistent phenomenon rather than an incidental observation.

during pretraining between representations of different modalities, when the multimodal model is exposed to the modal-specific distribution shift. We use t-SNE Van der Maaten & Hinton (2008) to visualize it in Fig. 1. As we can observe, when no distribution shift is present, the representations of different modalities extracted by multimodal models Gong et al. (2023) exhibit an approximately symmetric geometry. This is because, when two modality-specific encoders are pretrained with a contrastive loss Chen et al. (2020b); Li et al. (2022a), both encoders are optimized to organize their representations according to the same alignment and discrimination principle. Consequently, the resulting representations tend to exhibit comparable compactness. However, under the modal-specific distribution shift, the corrupted modality exhibits increased compactness in its representation geometry, whereas the other modality remains relatively stable, thus resulting in geometric symmetry non-conservation. The experimental results in Fig. 2 reveal a strong correlation between geometric symmetry non-conservation and the performance degradation of multimodal models under the modal-specific distribution shift. Moreover, we investigate the cause of geometric symmetry non-conservation by providing an intuitive explanation from the perspective of a probabilistic graphical model Shalev-Shwartz & Ben-David (2014) and a rigorous theoretical derivation.

Motivated by two key findings, we seek to address the aforementioned issue by *reforming* geometric symmetry, thereby improving the robustness of multimodal models to the modal-specific distribution shift. (1) Numerous works Huang et al. (2025); Klebe et al. (2023); Chen et al. (2025a) have shown that preserving the geometric structure formed during pretraining benefits downstream performance. (2) Our exploratory results in Fig. 2 show that, under the modal-specific distribution shift, the corruption degree of geometric symmetry is positively correlated with the extent of the model's performance degradation. To this end, we propose an effective approach ***ManiGest***, short for ***Mani****fold-constraint* ***G****eometry* ***Rest****oration*, to mitigate the impact of the modal-specific distribution shift on geometric symmetry via manifold-constrained techniques. Specifically, under the constraint of probabilistic Birkhoff polytope manifold, we aim to restore the geometric symmetry of representations across different modalities. However, a widely acknowledged multimodal learning assumption Tsai et al. (2021); Li et al. (2022b); Wang et al. (2025) is that the data of each modality contain both modality-consistent and modality-complementary task-dependent discriminative knowledge, which indicates that enforcing perfect geometric symmetry can sacrifice modality-

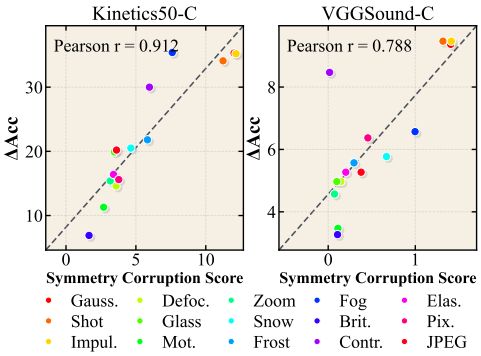

Figure 2: We apply 15 types of corruptions into the video modality on the Kinetics50 Kay et al. (2017) and VGGSound Chen et al. (2020a) datasets, and measure two metrics under each corruption: accuracy degradation (ΔAcc) and the degree of geometric symmetry disruption, quantified by the Symmetry Corruption Score introduced in **Appendix** A.4. We then compute the Pearson correlation coefficient between these two metrics. The results show a strong positive correlation between these two quantities, with coefficients of 0.912 and 0.788 on two datasets ("C" means the dataset is corrupted).

specific discriminative knowledge. Therefore, we characterize intra-modal geometry as elements on the symmetric positive definite (SPD) manifold, and preserve such geometry via the affine-invariant Riemannian distance on the SPD manifold. The empirical results demonstrate the effectiveness of our proposed method as a plug-and-play module.

Our contributions can be summarized as follows: (**1**) Through empirical evidence, we reveal an overlooked yet persistent issue that hinders the development of MTTA, which we term geometric symmetry non-conservation. We further investigate its cause through graphical explanation and the theoretical deduction. (**2**) We propose ManiGest to address the issue of geometric symmetry non-conservation, thus improving the robustness of multimodal models against the modal-specific distribution shift. (**3**) We conduct extensive experiments to demonstrate the effectiveness of ManiGest as a plug-and-play module to the modal-specific distribution shift on downstream tasks.

2 Related Work

Test-time adaptation seeks to adapt a pretrained model to a target domain without accessing the source data. The primary challenge is to rely solely on unlabeled target data to mitigate distribution shifts between source and target domains. Early studies primarily focus on unimodal settings. For example, Tent Wang et al. (2021) adapts the normalization layers by minimizing the entropy of model predictions, thereby reducing cross-domain distribution discrepancies. EATA Niu et al. (2022) and SAR Niu et al. (2023) incorporate sample selection strategies to filter unreliable samples and emphasize informative ones, leading to more stable and effective adaptation.

The effectiveness of UTTA methods often deteriorates in multimodal scenarios due to the increased complexity of distribution shifts, prompting the development of various MTTA approaches. For instance, MM-TTA Shin et al. (2022) alleviates the distribution shift through complementary pseudo-labeling across modalities, but its robustness remains contingent on the quality of the generated pseudo-labels; READ Yang et al. (2024) proposes a more robust approach by enhancing the impact of confident predictions and reducing the influence of noisy predictions. TSA Chen et al. (2025b) further addresses the modal-specific distribution shift problem through a selective adaptation framework designed to handle shifts in the affected modality. More recently, ABPEM Zhao et al. (2025) enhances multimodal TTA by reducing modality misalignment under distribution shifts via attention bootstrapping and principal entropy minimization.

Despite these advances, existing MTTA methods overlook a persistent issue at the representational level, i.e., geometric symmetry non-conservation. Correspondingly, we propose a novel manifold-constrained geometry restoration technique to address this shortcoming.

3 The Investigation on Geometric Symmetry Non-Conservation

Following the setup of representative baselines Yang et al. (2024); Chen et al. (2025b); Zhao et al. (2025), we take audio and video as an example to define the notation. Let $\mathcal{M} = \{f^a, f^v, g, h\}$ denote the multimodal model pretrained on the source domain, where f^a and f^v are modality-specific encoders, g is a multimodal attention fusion module Vaswani et al. (2017) parameterized by W_Q, W_K, W_V, and h is the classification head. Given an input multimodal sample $\boldsymbol{x} = (x^a, x^v)$, the prediction logits are computed as $\mathcal{M}(\boldsymbol{x}) = h\left(g\left[f^a(x^a); f^v(x^v)\right]\right)$.

We provide an intuitive explanation from the perspective of probabilistic graphical model Shalev-Shwartz & Ben-David (2014); Bishop & Nasrabadi (2006). Let $m \in \{a, v\}$ denote an arbitrary modality, and let $x^m_{y_i}$ and $x^m_{y_j}$ be samples from modality m with labels y_i and y_j ($y_i \neq y_j$), respectively. The ideal data generation process is illustrated in Fig. 3(a). Specifically, the label determines the generation of the observed data, and thus we have $y_i \to x^m_{y_i}$ and $y_j \to x^m_{y_j}$. $x^m_{y_i}$ is independent of $x^m_{y_j}$ due to $y_i \neq y_j$. However, when the m-th modality is corrupted and consequently undergoes a distribution

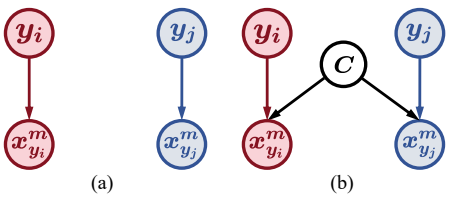

Figure 3: (a) illustrates the data generation process under the ideal condition, whereas (b) depicts the data generation process under distribution shift.

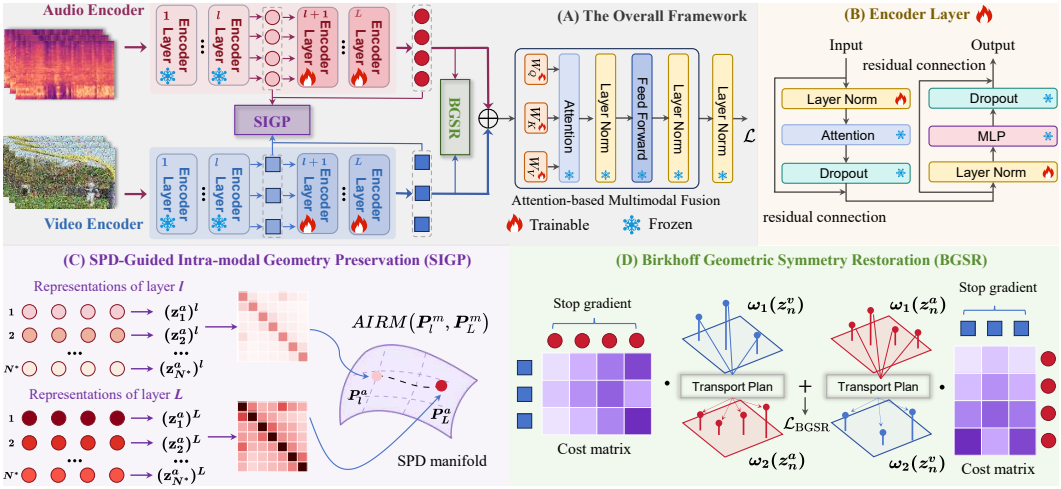

Figure 4: Overview of the proposed ManiGest framework. (A) illustrates the overall architecture of ManiGest. (B) illustrates the architecture of a trainable encoder layer, while (C) and (D) describe the SPD-Guided Intra-modal Geometry Preservation module and the Birkhoff Geometric Symmetry Restoration module, respectively. Due to space limitations, we use the audio modality as a representative example to illustrate the SPD-Guided Intra-modal Geometry Preservation module.

shift, the data generation process of all samples within modality m is interfered with by the common corruption c. In this case, the data generation process is depicted in Fig. 3(b). Specifically, $x_{y_i}^m$ is determined by y_i, c and $x_{y_j}^m$ is determined by y_j, c, i.e., $y_i \to x_{y_i}^m \leftarrow c$ and $y_j \to x_{y_j}^m \leftarrow c$. According to the independence relations implied by the probabilistic graphical model, $x_{y_i}^m$ and $x_{y_j}^m$ become dependent due to a common cause (or corruption). Overall, the corruption perturbs the data-generating process and renders any two originally independent samples in this modality correlated, making the features of all samples in this modality become geometrically more compact.

Theorem 3.1 provides a formal justification for the intuitive conclusion derived from the perspective of probabilistic graphical model, and we prove Theorem 3.1 in **Appendix** A.5.

Theorem 3.1. *Let $m \in \{a, v\}$ denote an arbitrary modality, $x_{y_i}^m$ and $x_{y_j}^m$ be samples drawn from $p(x^m \mid y_i)$ and $p(x^m \mid y_j)$, where $y_i \neq y_j$. Given $z_{y_i}^m = f^m(x_{y_i}^m)$, $z_{y_j}^m = f^m(x_{y_j}^m)$ and corresponding representations $\hat{z}_{y_i}^m, \hat{z}_{y_j}^m$ perturbed by distribution shift. For any $\delta > 0$, with high probability at least $1 - \delta$, the following inequality holds:*

$$\mathbb{E}\left(\left\langle \hat{z}_{y_i}^m, \hat{z}_{y_j}^m \right\rangle\right) \geq \mathbb{E}\left(\left\langle z_{y_i}^m, z_{y_j}^m \right\rangle\right), \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the Cosine similarity function.

Building upon Theorem 3.1, we can explain the cause of geometric symmetry non-conservation with sufficient theoretical support. Due to contrastive learning during pretraining, representations from different modalities tend to exhibit approximate geometric symmetry. Under the modal-specific distribution shift, the similarity between representations of any two distinct classes increases in the corrupted modality, making its representations geometrically more compact, while the uncorrupted modality remains relatively stable, thereby leading to geometric symmetry non-conservation.

Geometric symmetry non-conservation disrupts the structural correspondence between modality-specific representations, thereby degrading discriminability and multimodal fusion. Empirically, our experimental results in Fig. 2 reveal a strong correlation between geometric symmetry non-conservation and the performance degradation of multimodal models under the modal-specific distribution shift. To this end, we propose ManiGest to address this issue, thus improving the robustness of multimodal models against the modal-specific distribution shift.

4 METHOD

The framework of ManiGest is illustrated in Fig. 4. In practice, it is generally unknown which modality is affected by the distribution shift in practical scenarios. Therefore, throughout this section, we do not explicitly distinguish between corrupted and uncorrupted modalities in our notation.

4.1 BIRKHOFF GEOMETRIC SYMMETRY RESTORATION

Given a minibatch of input multimodal samples $\{\boldsymbol{x}_n = (x_n^a, x_n^v)\}_{n=1}^{N^*}$, the corresponding representations of the m-th ($(m \in \{a, v\})$) modality from l-th layer are obtained as

$$(\boldsymbol{z}_n^m)^l = f_l^m(x_n^m) = [(z_{n,1}^m)^l, (z_{n,2}^m)^l, \cdots, (z_{n,T_m}^m)^l]^\top \in \mathbb{R}^{T_m \times d},$$

where $l \in \{1, \ldots, L\}$, N^* is the batch size, T_m is the token number of modality m, and d is the dimensionality of each token representation. For conciseness, we omit the superscript associated with the final-layer (L-th layer) representations, i.e., we have $(\boldsymbol{z}_n^m)^L = \boldsymbol{z}_n^m$ and the same convention applies to token-level representations.

InfoNCE-based contrastive learning used in the pretraining stage Chen et al. (2020b); Gong et al. (2023) can restore geometric symmetry to a certain extent. Nevertheless, InfoNCE typically computes sample-level similarities by applying mean pooling to all token representations of each sample, which introduces two limitations: (1) It fails to account for the varying importance of different tokens; (2) It collapses token-level representations into a single global vector, thereby losing fine-grained geometric structure Chen et al. (2022).

To address these issues, drawing inspiration from Chen et al. (2022); Zhou et al., we construct a token-level transport plan between $\boldsymbol{z}_n^a$ and $\boldsymbol{z}_n^v$ under the constraint of probabilistic Birkhoff polytope manifold Beck & Pixton (2003); Xie et al. (2025) (Definition 4.1). The resulting Wasserstein distance Villani (2009) serves as a fine-grained alternative to the cosine similarity computed from mean-pooled sample-level representations.

Definition 4.1. *Let $p_1 \in \Delta_M$ and $p_2 \in \Delta_N$ be two probability distributions with strictly positive entries. The probabilistic Birkhoff polytope manifold is a generalization of Birkhoff polytope manifold Beck & Pixton (2003) and can be defined as:*

$$\mathcal{B}^{M \times N}(p_1, p_2) = \{\boldsymbol{T} \in \mathbb{R}_+^{M \times N} \mid \boldsymbol{T} \cdot \mathbf{1}_N = p_1, \boldsymbol{T} \cdot \mathbf{1}_M^\top = p_2\}, \tag{2}$$

(***Remark.*** The standard Birkhoff polytope manifold consists of non-negative matrices whose row and column sums are all equal to 1, whereas $\mathcal{B}^{M \times N}(p_1, p_2)$ consists of non-negative matrices whose row and column sums follow probability distributions.)

However, under the modal-specific distribution shift, only one modality is corrupted, whereas the other remains reliable. Directly minimizing the Wasserstein distance between the two modalities may therefore propagate the perturbation from the corrupted modality to the uncorrupted one, forcing reliable representations to align with the corrupted representations. To avoid this undesirable effect, we propose a unilateral gradient-truncated Wasserstein distance under the probabilistic Birkhoff polytope constraint, inspired by the stop-gradient strategy for preventing representation collapse Chen & He (2021); Grill et al. (2020).

Ideally, during the adaptation of the audio branch, the detached video features serve as fixed reference signals. Considering the varying quality of token representations, it is desirable to guide audio adaptation primarily with high-quality video tokens while assigning larger weights to low-quality audio tokens, thereby enabling more targeted correction of corrupted or unreliable audio token representations. The same principle applies symmetrically when adapting the video branch.

To facilitate this purpose, we provide two approaches to estimate the distribution of token representations corresponding to $\boldsymbol{z}_n^m$:

$$\mathcal{P}^1_{\boldsymbol{z}_n^m}(z) = \sum\nolimits_{t=1}^{T_m} \omega_1(z_{n,t}^m) \cdot \delta(z - z_{n,t}^m) \quad \text{and} \quad \mathcal{P}^2_{\boldsymbol{z}_n^m}(z) = \sum\nolimits_{t=1}^{T_m} \omega_2(z_{n,t}^m) \cdot \delta(z - z_{n,t}^m), \tag{3}$$

where $\delta(\cdot)$ is the Dirac function. Since entropy reflects uncertainty Ash (1990) (higher entropy indicating greater uncertainty), we introduce two entropy-based weight functions $\omega_1(\cdot), \omega_2(\cdot)$ to

measure the uncertainty and certainty of the tokens, respectively:

$$\omega_1(z_{n,t}^m) = \frac{\exp\left(\mathcal{H}\left(z_{n,t}^m\right)\right)}{\sum_{j=1}^{T_m}\exp\left(\mathcal{H}\left(z_{n,j}^m\right)\right)} \quad \text{and} \quad \omega_2(z_{n,t}^m) = \frac{\exp\left(\log K - \mathcal{H}\left(z_{n,t}^m\right)\right)}{\sum_{j=1}^{T_m}\exp\left(\log K - \mathcal{H}\left(z_{n,j}^m\right)\right)}, \tag{4}$$

where $h(z_{n,t}^m) \in \mathbb{R}^K$ (K is the number of categories) is the prediction of $z_{n,t}^m$, and $\mathcal{H}\left(z_{n,t}^m\right) = \sum_{k=1}^{K} -h(z_{n,t}^m)_k \log h(z_{n,t}^m)_k$ is the entropy of $z_{n,t}^m$.

Based on these two empirical distributions, we define the unilateral gradient-truncated Wasserstein distance from the audio modality to the detached video modality as

$$\mathcal{W}\left(\mathcal{P}^1_{\boldsymbol{z}_n^a}(z), sg\left(\mathcal{P}^2_{\boldsymbol{z}_n^v}(z)\right)\right) = \sum_{t_1}\sum_{t_2} \boldsymbol{T}_{t_1t_2}\cdot \boldsymbol{C}_{t_1t_2}(\boldsymbol{z}_n^a, \boldsymbol{z}_n^v), \text{ s.t. } \boldsymbol{T} \in \mathcal{B}^{T_a\times T_v}\left(\omega_1(z_n^a), \omega_2(z_n^v)\right).$$

$sg(\cdot)$ denote stop gradient, $\boldsymbol{C}(\boldsymbol{z}_n^a, \boldsymbol{z}_n^v)$ is the cost matrix, and we have $\boldsymbol{C}_{t_1t_2}(\boldsymbol{z}_n^a, \boldsymbol{z}_n^v) = 1 - < z_{n,t_1}^a, z_{n,t_2}^v >$. The optimization objective for geometric symmetry restoration can be formalized as

$$\mathcal{L}_{\text{BGSR}} = \frac{1}{N^*}\sum\nolimits_{n=1}^{N^*}\left(\mathcal{W}\left(\mathcal{P}^1_{\boldsymbol{z}_n^a}(z), sg\left(\mathcal{P}^2_{\boldsymbol{z}_n^v}(z)\right)\right) + \mathcal{W}\left(\mathcal{P}^1_{\boldsymbol{z}_n^v}(z), sg\left(\mathcal{P}^2_{\boldsymbol{z}_n^a}(z)\right)\right)\right). \tag{5}$$

4.2 SPD-Guided Intra-modal Geometry Preservation

A widely adopted assumption in multimodal learning Tsai et al. (2021); Li et al. (2022b); Wang et al. (2025) is that the task-dependent information contained in each unimodal observation consists of both modality-consistent and modality-complementary discriminative knowledge. However, imposing overly strong geometric symmetry undesirably homogenizes intra-modality geometry of different modalities. As a result, multimodal models may become biased toward modality-consistent information and fail to fully exploit modality-specific discriminative cues. To avoid this undesirable effect, we introduce an additional regularization term to preserve the intra-modal geometry entailed by the representations of each modality, while performing geometric symmetry restoration.

As discussed in **Section** 1, preserving the geometric structure formed during pretraining can benefit downstream performance. Moreover, prior studies Huang et al. (2024); Qiang et al. (2025); Wang et al. (2024); Yosinski et al. (2014) have shown that shallow-layer features generally retain more pretrained knowledge than deep-layer features. Motivated by these observations, we capture the geometric structure of shallow-layer representations as a reference to preserve intra-modal geometry. The elements of SPD manifold is employed to capture the intra-modal geometry (Definition 4.2).

Definition 4.2. *The symmetric positive definite (SPD) manifold is defined as the set of all $N \times N$ real symmetric positive definite matrices:*

$$\mathcal{S}_{++}^N = \left\{\mathbf{X} \in \mathbb{R}^{N\times N} \mid \mathbf{X} = \mathbf{X}^\top, v^\top \mathbf{X} v > 0,\ \forall\, v \neq 0\right\}. \tag{6}$$

Specifically, given a minibatch of observations $\{x_n^m\}_{n=1}^{N^*}$ from an arbitrary modality $m \in \{a, v\}$, we can obtain the representations from the l-th layer $\{(\boldsymbol{z}_n^m)^l\}_{n=1}^{N^*}$. We obtain the global representation by averaging over the token representations $(\mathrm{z}_n^m)^l = \sum_{t=1}^{T_m}\omega_2\left((z_{n,t}^m)^l\right)\cdot(z_{n,t}^m)^l \in \mathbb{R}^d$.

Based on the global representations in the minibatch, we construct a kernel matrix $\mathbf{K}_l^m$ to encode the intra-modal geometry of modality-m representations at the l-th layer:

$$\boldsymbol{K}_l^m = \begin{pmatrix} K_l^m(1,1) & K_l^m(1,2) & \cdots & K_l^m(1,N^*) \\ K_l^m(2,1) & K_l^m(2,2) & \cdots & K_l^m(2,N^*) \\ \vdots & \vdots & \ddots & \vdots \\ K_l^m(N^*,1) & K_l^m(N^*,2) & \cdots & K_l^m(N^*,N^*) \end{pmatrix}. \tag{7}$$

Each entry measures the pairwise similarity between the n_1-th and n_2-th samples in the kernel-induced feature space:

$$K_l^m(n_1,n_2) = \exp\left(\frac{< (\mathrm{z}_{n_1}^m)^l,\ (\mathrm{z}_{n_2}^m)^l >}{\tau}\right), \tag{8}$$

where $n_1, n_2 \in [1, N^*]$ and τ is a temperature parameter. This kernel nonlinearly amplifies pairwise similarities, allowing the resulting kernel matrix to capture richer relational structures than a linear

Table 1: Prediction accuracies (in %) on Kinetics50-C under corrupted video modality. The accuracy values for each corruption type are presented, along with the average accuracy. "*" indicates that the result is reproduced by ourselves, and the following table is the same.

Models	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	
Source	46.8	48	46.9	67.5	62.2	70.8	66.7	61.6	60.3	46.7	75.2	52.1	65.7	66.5	61.9	59.9
Tent	46.3	47	46.3	67.2	62.5	71	67.6	63.1	61.1	34.9	75.4	51.6	66.8	67.2	62.7	59.4
EATA	46.8	47.6	47.1	67.2	62.7	70.6	67.2	62.3	60.9	46.7	75.2	52.4	65.9	66.8	62.5	60.1
SAR	46.7	47.4	46.8	67	61.9	70.4	66.4	61.8	60.6	46	75.2	52.1	65.7	66.4	62	59.8
MMT	46.2	46.6	46.1	58.8	55.7	62.4	61.7	52.6	54.4	48.5	69.3	49.3	57.6	56.4	54.5	54.5
READ	49.4	49.7	49.0	68.0	65.1	71.2	69.0	64.5	64.4	57.4	75.5	53.6	68.3	68.0	65.1	62.5
+ ManiGest	51.9	52.0	51.7	70.0	67.4	71.7	70.2	66.3	67.7	64.5	76.4	56.6	70.6	70.0	67.6	65.0
Δ	+2.5	+2.3	+2.7	+2.0	+2.3	+0.5	+1.2	+1.8	+3.3	+7.1	+0.9	+3.0	+2.3	+2.0	+2.5	+2.5
TSA	52.6	52.3	52.0	68.7	68.0	70.7	68.8	65.2	66.6	64.3	74.6	57.4	70.5	69.0	66.2	64.5
+ ManiGest	54.8	56.5	56.0	69.4	69.3	71.4	69.2	66.5	67.7	65.6	76.2	59.3	71.9	71.7	68.9	66.3
Δ	+2.2	+4.2	+4.0	+0.7	+1.3	+0.7	+0.4	+1.3	+1.1	+1.3	+1.6	+1.9	+1.4	+2.7	+2.7	+1.8
ABPEM*	50.3	50.9	50.6	69.5	69.2	72.3	71.7	66.0	67.1	65.7	75.4	56.2	71.6	71.2	68.2	65.0
+ ManiGest	53.4	55.9	54.8	70.4	69.8	73.6	70.7	66.2	68.0	66.2	76.9	58.2	71.5	72.0	69.7	66.5
Δ	+3.0	+5.0	+4.2	+0.9	+0.6	+1.3	-1.0	+0.2	+0.9	+0.5	+1.6	+2.0	-0.1	+0.8	+1.6	+1.4

Table 2: Prediction accuracies (in %) on Kinetics50-C (left) and VGGSound-C (right) under corrupted audio modality. Accuracies are reported for each corruption type, together with the average accuracy across all corruptions.

Models	Noise			Weather			Avg.	Noise			Weather			Avg.
	Gauss.	Traff.	Crowd.	Rain	Thund.	Wind		Gauss.	Traff.	Crowd.	Rain	Thund.	Wind	
Source	73.7	65.5	67.9	70.3	67.9	70.3	69.3	37.0	25.5	16.8	21.6	27.3	25.5	25.6
Tent	73.9	67.4	69.2	70.4	66.5	70.5	69.6	10.6	2.6	1.8	2.8	5.3	4.1	4.5
EATA	73.7	66.1	68.5	70.3	67.9	70.1	69.4	39.2	26.1	22.9	26.0	31.7	30.4	29.4
SAR	73.7	65.4	68.2	69.9	67.2	70.2	69.1	37.4	9.5	11.0	12.1	26.8	23.7	20.1
MMT	70.8	69.2	68.5	69.0	69.8	68.5	69.4	14.1	5.2	6.4	9.8	8.6	4.5	7.6
READ	74.1	69.0	69.7	71.1	71.8	70.7	71.1	40.4	28.9	26.6	30.9	36.7	30.6	32.4
+ ManiGest	75.2	71.5	71.5	71.8	74.2	72.4	72.7	41.1	31.7	29.4	32.0	38.3	31.4	34.0
Δ	+1.1	+2.5	+1.8	+0.7	+2.4	+1.7	+1.6	+0.7	+2.8	+2.8	+1.1	+1.6	+0.8	+1.6
TSA	74.5	69.6	70.5	71.4	72.0	71.0	71.5	41.5	31.8	30.9	32.6	38.9	32.6	34.7
+ ManiGest	75.0	70.0	71.4	71.6	72.3	71.2	71.9	42.0	33.5	32.2	32.4	39.8	33.7	35.6
Δ	+0.5	+0.4	+0.9	+0.2	+0.3	+0.2	+0.4	+0.5	+1.7	+1.3	-0.2	+0.9	+1.1	+0.9
ABPEM*	74.2	71.2	71.3	71.3	73.8	71.4	72.2	36.0	24.4	24.0	23.3	31.1	23.8	27.1
+ ManiGest	73.9	71.3	72.1	71.7	73.1	71.9	72.3	40.8	32.7	31.8	32.9	40.1	32.7	35.1
Δ	-0.3	+0.1	+0.8	+0.4	-0.8	+0.5	+0.1	+4.8	+8.3	+7.8	+9.6	+9.0	+8.8	+8.1

Gram matrix. The temperature parameter τ further controls the sensitivity of the induced geometry to similarity variations. To ensure strict positive definiteness and numerical stability, we regularize the similarity matrix as $\mathbf{P}_l^m = \mathbf{K}_l^m + \epsilon\mathbf{I}$, where $\epsilon > 0$ is a small constant and $\mathbf{I}$ is the $N^* \times N^*$ identity matrix. Obviously, we have $\mathbf{P}_l^m \in \mathcal{S}_{++}^{N^*}$. Since $\mathbf{P}_l^m$ and $\mathbf{P}_L^m$ encode batch-wise relational structures of the same modality and lie on the same SPD manifold $\mathcal{S}_{++}^{N^*}$, we measure their discrepancy in a manner consistent the intrinsic geometry of the SPD manifold. Accordingly, we employ the affine-invariant Riemannian metric (AIRM) Pennec et al. (2006), which induces the geodesic distance on $\mathcal{S}_{++}^{N^*}$:

$$\mathcal{L}_{\text{SPD}}^m = \text{AIRM}(\mathbf{P}_l^m, \mathbf{P}_L^m) = \left\| \log\left((\mathbf{P}_l^m)^{-1/2} \mathbf{P}_L^m (\mathbf{P}_l^m)^{-1/2} \right) \right\|_F = \sqrt{\sum_{k=1}^{N^*} (\log \lambda_k^m)^2}, \quad (9)$$

where $\{\lambda_k^m\}_{k=1}^{N^*}$ denote the eigenvalues of $(\mathbf{P}_l^m)^{-1/2} \mathbf{P}_L^m (\mathbf{P}_l^m)^{-1/2}$ and $\|\cdot\|_F$ denotes the Frobenius norm. The SPD-guided intra-modal geometry preservation loss is defined over both modalities as

$$\mathcal{L}_{\text{SIGP}} = \mathcal{L}_{\text{SPD}}^a + \mathcal{L}_{\text{SPD}}^v. \quad (10)$$

Given the objective function $\mathcal{L}$ of an arbitrary MTTA baseline, such as READ Yang et al. (2024) or ABPEM Zhao et al. (2025), ManiGest can be integrated as a plug-and-play module by optimizing the following objective:

$$\mathcal{L}_{\text{ManiGest}} = \mathcal{L} + \alpha \cdot \mathcal{L}_{\text{BGSR}} + \beta \cdot \mathcal{L}_{\text{SIGP}}, \quad (11)$$

where α and β are the hyper-parameters to control the balance of $\mathcal{L}_{\text{BGSR}}$ and $\mathcal{L}_{\text{SIGP}}$, respectively. The training pipeline is depicted by Algorithm 1 in **Appendix** A.6

Table 3: Prediction accuracies (in %) on VGGSound-C under corrupted video modality. The accuracy values for each corruption type are presented, along with the average accuracy.

Models	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	
Source	52.8	52.7	52.7	57.2	57.2	58.7	57.6	56.4	56.6	55.6	58.9	53.7	56.9	55.8	56.9	56.0
Tent	52.7	52.7	52.7	56.7	56.5	57.9	57.2	55.9	56.3	56.3	58.4	54.0	57.4	56.2	56.7	55.8
EATA	53.0	52.8	53	57.2	57.1	58.6	57.8	56.3	56.8	56.4	59.0	54.1	57.4	56.1	57.0	56.2
SAR	52.9	52.8	52.9	57.2	57.1	58.6	57.6	56.3	56.7	55.9	58.9	54.0	57.0	56.0	57.0	56.1
MMT	7.1	7.3	7.3	44.8	41.5	48.0	45.5	27.4	23.5	30.5	46.3	24.0	43.0	40.7	45.7	32.0
READ	53.6	53.6	53.5	57.9	57.7	59.4	58.8	57.2	57.8	55.0	59.9	55.2	58.6	57.1	57.9	56.9
+ ManiGest	54.5	54.6	54.4	59.2	58.7	60.1	59.5	58.0	58.3	58.3	60.4	55.9	59.2	58.5	58.9	57.9
Δ	+0.9	+1.0	+0.9	+1.3	+1.0	+0.7	+0.7	+0.8	+0.5	+3.3	+0.5	+0.7	+0.6	+1.4	+1.0	+1.0
TSA	53.9	53.9	54.0	57.6	58.0	59.0	58.6	56.9	57.0	56.6	59.8	54.7	58.6	56.7	57.9	56.9
+ ManiGest	54.2	54.2	54.4	58.3	58.4	59.5	58.8	57.2	57.7	57.5	60.0	55.7	58.8	58.2	58.4	57.4
Δ	+0.3	+0.3	+0.4	+0.7	+0.4	+0.5	+0.2	+0.3	+0.7	+0.9	+0.2	+1.0	+0.2	+1.5	+0.5	+0.5
ABPEM*	53.9	54.0	53.8	58.2	58.3	59.5	59.3	57.1	57.9	58.1	60.3	55.9	59.3	57.5	58.3	57.4
+ ManiGest	54.7	54.7	54.5	58.5	58.7	60	59.4	57.6	58.3	58.6	60.5	56.2	59.5	58.7	58.8	57.9
Δ	+0.8	+0.7	+0.7	+0.3	+0.4	+0.5	+0.1	+0.5	+0.4	+0.5	+0.2	+0.3	+0.2	+1.2	+0.5	+0.5

Table 4: The results of component ablation on Kinetics50-C video corruptions. Δ denotes the improvement over baseline.

Method	BGSR	SIGP	Avg.	Δ
READ			62.5	–
+ ManiGest	✓		64.8	+2.3
+ ManiGest		✓	64.5	+2.0
+ ManiGest	✓	✓	**65.0**	**+2.5**
ABPEM			65.0	–
+ ManiGest	✓		66.2	+1.2
+ ManiGest		✓	66.3	+1.3
+ ManiGest	✓	✓	**66.5**	**+1.5**

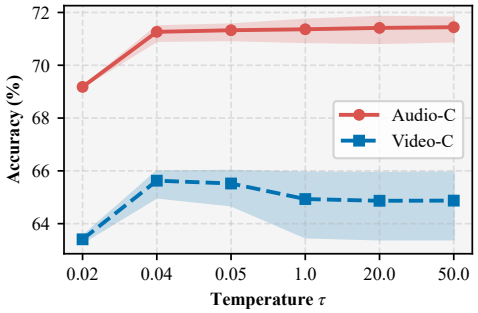

Figure 5: Sensitivity analysis with respect to the temperature parameter τ.

5 Experiments

Due to space constraints, we defer the descriptions of the datasets, baselines, evaluation protocol, and implementation details to **Appendix** A.2. The code for this work is provided in the submitted supplementary materials.

5.1 Performance Comparison

Tables 1, 2, and 3 report the performance under modality-specific corruptions. Overall, ManiGest consistently improves competitive multimodal TTA baselines across different datasets, corrupted modalities, and corruption types, demonstrating its effectiveness as a plug-and-play module. In addition to average improvements, ManiGest also shows stable gains on several corruption types across different settings. For example, under video corruptions, positive gains are repeatedly observed on noise corruptions such as *Gauss.*, *Shot*, and *Impul.*, as well as digital corruptions such as *Pix.* and *JPEG*. Under audio corruptions, ManiGest also brings consistent improvements on *Traff.*, *Crowd.*, and *Wind.* across datasets and base methods. These results suggest that ManiGest complements existing adaptation objectives by introducing geometry restoration, thereby improving the robustness of MTTA under the modal-specific distribution shift.

5.2 Ablation Study

Component Ablation. We conduct an ablation study to examine the individual contributions of two proposed principal components: Birkhoff Geometric Symmetry Restoration (BGSR) and SPD-Guided Intra-modal Geometry Preservation (SIGP). As shown in Table 4, each component improves the corresponding baselines, and their combination achieves the best average performance. These results indicate that BGSR and SIGP are both effective and complementary for robust multimodal test-time adaptation. Detailed results for each corruption are provided in **Appendix** A.7.

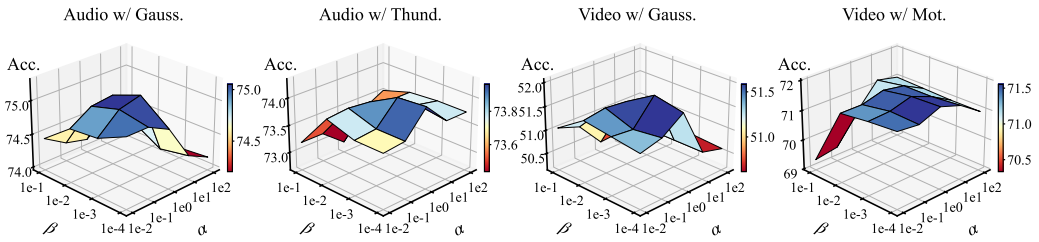

Figure 6: Parameter sensitivity analysis of READ+ManiGest on Kinetics50-C with respect to the balancing coefficients α and β. We show four representative cases covering audio and video corruptions; the complete results are provided in the **Appendix** A.7.

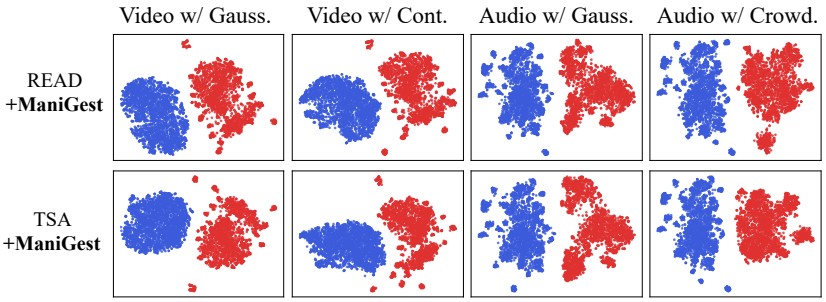

Figure 7: t-SNE Van der Maaten & Hinton (2008) visualization of adapted multimodal representations for representative different modality corruption cases. Columns denote corruption settings, and rows denote ManiGest incorporated into different baselines. Blue and red points indicate video and audio representations.

Hyperparameter Sensitivity. We further analyze the sensitivity of ManiGest to its key hyperparameters, including the temperature parameter τ and the balancing coefficients α and β of the two auxiliary losses. For the temperature parameter τ, we evaluate TSA+ManiGest under both audio and video corruptions. As shown in Fig. 5, the performance improves noticeably when τ increases from 0.02 to 0.04, after which the average accuracy remains relatively stable. When τ is further increased, the average accuracy remains broadly stable, and some larger values may even yield slightly higher results. However, the enlarged shaded region indicates that performance fluctuations become more pronounced, especially under Video-C, suggesting reduced stability across corruption types at larger temperature values. As shown in Fig. 6, we perform a grid search over α and β on Kinetics50-C under READ+ManiGest, and report the classification accuracy for representative audio and video corruptions. Across different corrupted modalities and corruption types, the accuracy exhibits mild variations under diverse parameter combinations. This suggests that the two loss terms can be balanced within a relatively flexible range. Overall, these results indicate that ManiGest is reasonably robust to both the temperature parameter and the loss balancing coefficients.

Visualization of Restored Representations. To qualitatively examine whether ManiGest helps restore geometric symmetry under the modal-specific distribution shift, we visualize the representations under three corruptions (*Gauss.*, *Cont*, and *Crowd.*) in Fig. 7. Compacted to the visualization in Fig. 1, the representations from different modalities exhibit more balanced compactness With ManiGest integrated, qualitatively suggesting that ManiGest helps restore geometric symmetry.

6 Conclusion

In this paper, we identify an overlooked yet persistent issue in MTTA, termed *geometric symmetry non-conservation*. We provide an intuitive explanation and a theoretical derivation to reveal its cause, and further show that such non-conservation is closely associated with the performance degradation of multimodal models under modal-specific distribution shifts. To address this issue, we propose ManiGest, a manifold-constrained geometry restoration framework. Extensive experiments on multiple benchmarks under diverse distribution shifts demonstrate the effectiveness and plug-and-play applicability of ManiGest.

AI USE STATEMENT

In this work, we used generative AI tools solely for language editing and polishing to improve the clarity, readability, and grammatical correctness of the manuscript. Generative AI tools were not used for research ideation, methodology development, experimental design, data analysis, result interpretation, or the generation of scientific claims. All AI-assisted revisions were carefully reviewed by the authors to ensure that they accurately preserved the intended meaning and scientific content. We take full responsibility for the final content of this work, including all text revised with the assistance of generative AI.

REFERENCES

Robert B Ash. *Information theory*. Courier Corporation, 1990.

Matthias Beck and Dennis Pixton. The ehrhart polynomial of the birkhoff polytope. *Discrete & Computational Geometry*, 30(4):623–637, 2003.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Dexia Chen, Qianjie Zhu, Weibing Li, Yue Yu, Tong Zhang, and Ruixuan Wang. Preserve and sculpt: Manifold-aligned fine-tuning of vision-language models for few-shot learning. *arXiv preprint arXiv:2508.12877*, 2025a.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020a.

Mingcai Chen, Baoming Zhang, Zongbo Han, Wenyu Jiang, Yanmeng Wang, Shuai Feng, Yuntao Du, and Bingkun Bao. Test-time selective adaptation for uni-modal distribution shift in multi-modal data. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025b. URL https://proceedings.mlr.press/v267/chen25ch.html.

Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. Benchmarking robustness of adaptation methods on pre-trained vision-language models. *Advances in Neural Information Processing Systems*, 36:51758–51777, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pp. 1597–1607. PMLR, 2020b. URL http://proceedings.mlr.press/v119/chen20j.html.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 15750–15758. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01549. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=QPtMRyk5rb.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3777–3786, 2025.

Xiaolong Huang, Qiankun Li, Xueran Li, and Xuesong Gao. One step learning, one step review. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12644–12652, 2024.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. URL http://arxiv.org/abs/1705.06950.

Yoonsik Kim, Jae Woong Soh, Gu Yong Park, and Nam Ik Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 3479–3489. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00354. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Kim_Transfer_Learning_From_Synthetic_to_Real-Noise_Denoising_With_Adaptive_Instance_CVPR_2020_paper.html.

Dustin Klebe, Tal Shnitzer, Mikhail Yurochkin, Leonid Karlinsky, and Justin Solomon. Gera: Label-efficient geometrically regularized alignment. *arXiv preprint arXiv:2310.00672*, 2023.

Jiangmeng Li, Wenwen Qiang, Yanan Zhang, Wenyi Mo, Changwen Zheng, Bing Su, and Hui Xiong. Metamask: Revisiting dimensional confounder for self-supervised learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022a. URL http://papers.nips.cc/paper_files/paper/2022/hash/fb575ab4d882a4c734641155a5f30911-Abstract-Conference.html.

Jiangmeng Li, Wenwen Qiang, Changwen Zheng, Bing Su, Farid Razzak, Ji-Rong Wen, and Hui Xiong. Modeling multiple views via implicitly preserving global consistency and local complementarity. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7220–7238, 2022b.

Jie Mu, Wei Wang, Wenqi Liu, Tiantian Yan, and Guanglu Wang. Multimodal large language model with lora fine-tuning for multimodal sentiment analysis. *ACM Trans. Intell. Syst. Technol.*, 16(6):139:1–139:23, 2025. doi: 10.1145/3709147. URL https://doi.org/10.1145/3709147.

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. *arXiv preprint arXiv:2302.12400*, 2023.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian Framework for Tensor Computing. *International Journal of Computer Vision*, 66(1):41–66, January 2006. ISSN 1573-1405. doi: 10.1007/s11263-005-3222-z. URL https://doi.org/10.1007/s11263-005-3222-z.

Wenwen Qiang, Zeen Song, Ziyin Gu, Jiangmeng Li, Changwen Zheng, Fuchun Sun, and Hui Xiong. On the generalization and causal explanation in self-supervised learning. *International Journal of Computer Vision*, 133(4):1727–1754, 2025.

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Sparsh Garg, In So Kweon, and Kuk-Jin Yoon. Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16928–16937, 2022.

Fei Song, Yi Li, Jiangmeng Li, Rui Wang, Changwen Zheng, Fanjiang Xu, and Hui Xiong. Ample: Supporting vision-language models via adaptive-debiased ensemble multi-prompt learning. *Int. J. Comput. Vis.*, 134(2):67, 2026. doi: 10.1007/S11263-025-02591-4. URL https://doi.org/10.1007/s11263-025-02591-4.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=-bdp_8Itjwp.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Cédric Villani. The wasserstein distances. In *Optimal transport: old and new*, pp. 93–111. Springer, 2009.

Chenyu Wang, Sharut Gupta, Xinyi Zhang, Sana Tonekaboni, Stefanie Jegelka, Tommi S. Jaakkola, and Caroline Uhler. An information criterion for controlled disentanglement of multimodal data. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=3n4RY25UWP.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, Francois Fleuret, and Pascal Frossard. Lines: Post-training layer scaling prevents forgetting and enhances model merging. *arXiv preprint arXiv:2410.17146*, 2024.

Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Kuai Yu, et al. mhc: Manifold-constrained hyper-connections. *arXiv preprint arXiv:2512.24880*, 2025.

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

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

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*, volume 39, pp. 22849–22857, 2025.

Haochen Zhou, Guofeng Ding, Mouxing Yang, Peng Hu, Yijie Lin, and Xi Peng. Uncover underlying correspondence for robust multi-view clustering. In *The Fourteenth International Conference on Learning Representations*.

A Appendix

A.1 Limitations

Although ManiGest provides a geometry-aware solution for improving MTTA, our theoretical analysis is limited to the modal-specific distribution shift modeled as additive noise. While this assumption offers a tractable and representative setting for studying how distribution shifts disrupt geometric symmetry, real-world shifts may involve more complex patterns. In addition, this work mainly focuses on the modal-specific distribution shift. However, in practical scenarios, distribution shifts may occur simultaneously across multiple or even all modalities, leading to more complicated cross-modal interactions and reliability variations. Thus, our analysis should be viewed as an initial step toward understanding geometric symmetry non-conservation, and extending it to more complex multimodal shift models remains an important direction for future work.

A.2 Experimental Settings

Datasets. We evaluate ManiGest on two audio-video multimodal datasets, Kinetics50-C and VGGSound-C, which are corrupted versions of Kinetics50 Kay et al. (2017) and VGGSound Chen et al. (2020a), respectively. We simulate the modal-specific distribution shift by applying 15 types of corruptions into the video modality and 6 types of corruptions into the audio modality. Detailed descriptions of these corruptions can be found in Yang et al. (2024). Accuracy is used as our evaluation metric.

Baselines. We compare the proposed method with a range of existing TTA methods, including UTTA methods such as TENT Wang et al. (2021), EATA Niu et al. (2022), and SAR Niu et al. (2023), as well as MTTA methods including MMT Shin et al. (2022), READ Yang et al. (2024), TSA Chen et al. (2025b), and ABPEM Zhao et al. (2025). For ABPEM, although we carefully follow the original implementation, the reproduced performance does not fully match the results reported in the original paper. We therefore report the reproduced results for comparison.

Evaluation Protocol. As a plug-and-play module, ManiGest is incorporated into representative MTTA baselines, including READ, TSA, and ABPEM. To ensure a fair comparison, we keep the original configuration of each baseline unchanged, including the learning rate, baseline-specific objectives, and other implementation settings. For each baseline, we compare the original method with its ManiGest-enhanced variant, where Δ denotes the performance improvement over the original baseline. This protocol provides a controlled evaluation of the compatibility and effectiveness of ManiGest across different MTTA methods and corruption settings.

Implementation Details Experiments are conducted on NVIDIA Tesla V100 and RTX 4090 GPUs. Specifically, the KS50-C reproduction and READ+ManiGest experiments are performed on an RTX 4090, while all other experiments are conducted on a V100. Each experiment is run on a single GPU. The hyperparameter settings for different ManiGest variants are summarized in Table 1, and all hyperparameters are selected in a label-free manner using the target test data. Unless otherwise specified, the baseline configurations follow their original implementations. For all experiments, we use the Adam optimizer with a learning rate of 10^{-4} and perform a single gradient-based adaptation step for each target batch. The backbone and classifier are kept frozen during adaptation, while only the LayerNorm parameters in the specified encoder blocks and the attention parameters in the multimodal fusion module are updated. For each corruption type, target samples are shuffled using a fixed random seed and partitioned into mini-batches. Each corruption stream is then processed

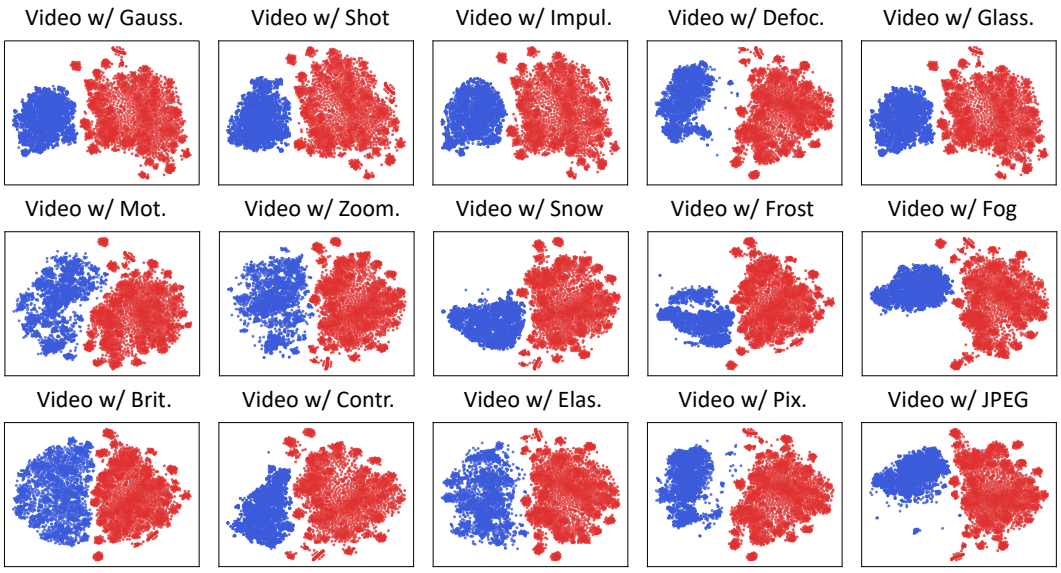

Figure 8: Extended visualization of geometric symmetry non-conservation under video corruptions. We introduce diverse corruptions into the video modality, including Gaussian, Shot, Impulse, Defocus blur, Glass blur, Motion blur, Zoom Blur, Snow, Frost, Fog, Brightness, Contrast, Elastic transformation, Pixelation, and JPEG compression. The blue and red points denote video and audio representations, respectively.

sequentially under a continual TTA protocol, where every target sample is encountered only once and the adapted model state is carried forward throughout the stream.

A.3 Extended Motivation Experiments

In this section, we provide extended motivation experiments to demonstrate that geometric symmetry non-conservation is a consistent phenomenon rather than an incidental observation. Specifically, on the Kinetics50 dataset, we introduce 15 corruptions for video modality and 6 corruptions for audio modality, respectively. The visualization results are depicted in Fig. 8 and Fig. 9. From the visualization results, we can observe that regardless of which modality is corrupted, the features of that modality become more compact, while the compactness of the features in the other modality remains relatively stable. This demonstrates that geometric symmetry non-conservation is a general phenomenon under the modal-specific distribution shift.

A.4 The Calculation of Symmetry Corruption Score

Before introducing the Symmetry Corruption Score, we first present the Symmetry Score. Specifically, let $\mathbf{Z}^a = \{\boldsymbol{z}_i^a\}_{i=1}^N$ and $\mathbf{Z}^v = \{\boldsymbol{z}_i^v\}_{i=1}^N$ denotes the paired representations from the audio and video modalities, respectively. We apply zero-mean normalization to each modality:

$$\tilde{\boldsymbol{z}}_i^a = \boldsymbol{z}_i^a - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{z}_j^a, \quad \tilde{\boldsymbol{z}}_i^v = \boldsymbol{z}_i^v - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{z}_j^v.$$

The symmetry score is then computed as the average ℓ_2 distance between paired normalized representations:

$$\mathrm{SymScore}(\mathbf{Z}^a, \mathbf{Z}^v) = \frac{1}{N}\sum_{i=1}^{N} \|\tilde{\boldsymbol{z}}_i^a - \tilde{\boldsymbol{z}}_i^v\|_2 .$$

Based on this score, we further define the Symmetry Corruption Score to quantify the extent to which a modal-specific distribution shift disrupts geometric symmetry. As an example, we construct a modal-specific distribution shift by introducing Gaussian noise into the video modality. Let $\mathbf{Z}^a = \{\boldsymbol{z}_i^a\}_{i=1}^N$ and $\mathbf{Z}^v = \{\boldsymbol{z}_i^v\}_{i=1}^N$ denote the paired representations from the uncorrupted audio and

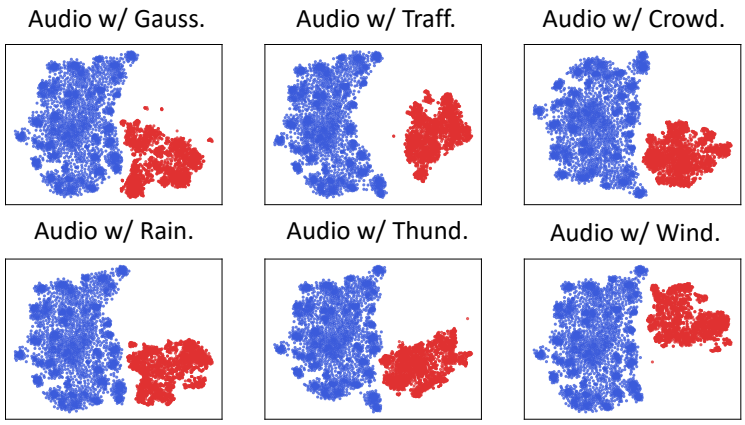

Figure 9: Extended visualization of geometric symmetry non-conservation under six audio corruptions. We introduce different types of corruption into the audio modality, including Gaussian, Traffic, Crowd, Rain, Thunder, and Wind. The blue and red points denote video and audio representations, respectively.

video modalities, respectively, and let $\hat{\mathbf{Z}}^v = \{\hat{\boldsymbol{z}}_i^v\}_{i=1}^N$ denote the perturbed video representations under Gaussian corruption. The Symmetry Corruption Score under this modal-specific distribution shift can be calculated by

$$\left|\mathrm{SymScore}(\mathbf{Z}^a, \mathbf{Z}^v) - \mathrm{SymScore}(\mathbf{Z}^a, \hat{\mathbf{Z}}^v)\right|. \tag{12}$$

A.5 Proof of Theorem 3.1

Proof. We prove the theorem under the zero-mean additive noise assumption Chen et al. (2025b); Kim et al. (2020). Specifically, the perturbed representations can be written as

$$\hat{z}_{y_i}^m = z_{y_i}^m + \varepsilon, \hat{z}_{y_j}^m = z_{y_j}^m + \varepsilon, \tag{13}$$

where $\mathbb{E}(\varepsilon) = \mathbf{0}$ and the covariance matrix of ε is Σ. Suppose that the category-conditioned probability distribution of $z_{y_i}^m$ and $z_{y_j}^m$ are $\mathcal{N}(\mu_i, \Sigma_i)$ and $\mathcal{N}(\mu_j, \Sigma_j)$, respectively. Our goal is to show that the expected similarity between the perturbed representations is no smaller than that between the original representations, i.e.,

$$\mathbb{E}\left(\langle \hat{z}_{y_i}^m, \hat{z}_{y_j}^m \rangle\right) \geq \mathbb{E}\left(\langle z_{y_i}^m, z_{y_j}^m) \rangle\right). \tag{14}$$

We first analyze the right-hand side of Equation equation 14, namely the expected similarity before distribution shift:

$$\begin{aligned}
\mathbb{E}\left(\langle z_{y_i}^m, z_{y_j}^m \rangle\right) &= \mathbb{E}\left(\sum_{k=1}^{d} [z_{y_i}^m]_k \cdot [z_{y_j}^m]_k\right) && \text{(Representations } z_{y_i}^m, z_{y_j}^m \text{ have been normalized)} \\
&= \sum_{k=1}^{d} \mathbb{E}\left([z_{y_i}^m]_k \cdot [z_{y_j}^m]_k\right) && \text{(Linearity of Expectation)} \\
&= \sum_{k=1}^{d} \mathbb{E}\left([z_{y_i}^m]_k\right) \cdot \mathbb{E}\left([z_{y_j}^m]_k\right) && (z_{y_i}^m, z_{y_j}^m \text{ are mutually independent}) \\
&= \left\langle \mathbb{E}(z_{y_i}^m), \mathbb{E}(z_{y_j}^m) \right\rangle \\
&= \langle \mu_i, \mu_j \rangle \\
&= \frac{\mu_i \cdot \mu_j}{\|\mu_i\| \cdot \|\mu_j\|}.
\end{aligned} \tag{15}$$

Next, we turn to the perturbed representations. To lower-bound their Cosine similarity, we first need to control the norm of the noise-perturbed representation. For this purpose, we recall Chebyshev's inequality.

Corollary A.1 (**Chebyshev's inequality**). *Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$, we have*

$$\mathbb{P}\left\{|X - \mu| \geq t\right\} \leq \frac{\sigma^2}{t^2}.$$

Consider the random variable $\|z_{y_i}^m + \boldsymbol{\varepsilon}\|^2$. Its expectation is given by

$$\mathbb{E}\left(\|z_{y_i}^m + \boldsymbol{\varepsilon}\|^2\right) = \mathbb{E}\left(\left(z_{y_i}^m + \boldsymbol{\varepsilon}\right)^\top \left(z_{y_i}^m + \boldsymbol{\varepsilon}\right)\right) = \|\mu_i\|^2 + \operatorname{trace}(\Sigma_i + \Sigma). \tag{16}$$

By applying Chebyshev's inequality, for any $\delta > 0$, we obtain

$$P\Big(\left|\frac{\|z_{y_i}^m + \boldsymbol{\varepsilon}\|^2}{\|\mu_i\|^2 + \operatorname{trace}(\Sigma_i + \Sigma)} - 1\right| \geq \delta\Big) \leq \frac{\operatorname{Var}\left(\frac{\|z_{y_i}^m + \boldsymbol{\varepsilon}\|^2}{\|\mu_i\|^2 + \operatorname{trace}(\Sigma_i + \Sigma)}\right)}{\delta^2}. \tag{17}$$

Therefore, with high probability, the following event holds:

$$\begin{aligned}\left|\frac{\|z_{y_i}^m + \boldsymbol{\varepsilon}\|^2}{\|\mu_i\|^2 + \operatorname{trace}(\Sigma_i + \Sigma)} - 1\right| < \delta \Longrightarrow -\delta < \frac{\|z_{y_i}^m + \boldsymbol{\varepsilon}\|^2}{\|\mu_i\|^2 + \operatorname{trace}(\Sigma_i + \Sigma)} - 1 < \delta \\ \Longrightarrow \|z_{y_i}^m + \boldsymbol{\varepsilon}\|^2 < (1+\delta)\left(\|\mu_i\|^2 + \operatorname{trace}(\Sigma_i + \Sigma)\right).\end{aligned} \tag{18}$$

Because (1) $\|\mu_i\|^2 + trace(\Sigma_i + \Sigma) \geq \|\mu_i\|^2$ and (2) function $(1+\delta)^x$ (where $\delta > 0$) is a monotonically increasing function of x with range $(0, +\infty)$, there must exist a constant $\mathrm{cons} > 0$ such that

$$\|\mu_i\|^2 + \operatorname{trace}(\Sigma_i + \Sigma) = (1+\delta)^{\mathrm{cons}} \cdot \|\mu_i\|^2. \tag{19}$$

Substitute Equation 19 into Equation 18:

$$\begin{aligned}\left|\frac{\|z_{y_i}^m + \boldsymbol{\varepsilon}\|^2}{\|\mu_i\|^2 + \operatorname{trace}(\Sigma_i + \Sigma)} - 1\right| < \delta &\Longrightarrow \|z_{y_i}^m + \boldsymbol{\varepsilon}\|^2 < (1+\delta)^{\mathrm{cons}+1} \cdot \|\mu_i\|^2 \\ &\Longrightarrow \frac{1}{\|z_{y_i}^m + \boldsymbol{\varepsilon}\|^2} > \frac{1}{(1+\delta)^{\mathrm{cons}+1} \cdot \|\mu_i\|^2} \\ &\Longrightarrow \frac{1}{\|z_{y_i}^m + \boldsymbol{\varepsilon}\|} > \frac{1}{\|\mu_i\|} \cdot (1+\delta)^{-1/2\cdot(\mathrm{cons}+1)} \\ &\Longrightarrow \frac{1}{\|z_{y_i}^m + \boldsymbol{\varepsilon}\|} > \frac{1}{\|\mu_i\|} \cdot \frac{1}{\left[1 - \frac{\delta}{2\cdot(\mathrm{cons}+1)}\right]}. \text{ (Taylor expansion)}\end{aligned} \tag{20}$$

We are now ready to compare the expected similarities before and after the distribution shift. Using the above norm bounds, we have

$$
\begin{aligned}
&\mathbb{E}\left(\langle z_{y_i}^m + \boldsymbol{\varepsilon}, z_{y_j}^m + \boldsymbol{\varepsilon}\rangle\right) \\
&= \mathbb{E}\left(\frac{(z_{y_i}^m + \boldsymbol{\varepsilon})^\top (z_{y_j}^m + \boldsymbol{\varepsilon})}{\|z_{y_i}^m + \boldsymbol{\varepsilon}\|\|z_{y_j}^m + \boldsymbol{\varepsilon}\|}\right) \\
&\geq \frac{\mathbb{E}\left((z_{y_i}^m + \boldsymbol{\varepsilon})^\top (z_{y_j}^m + \boldsymbol{\varepsilon})\right)}{(1 - \frac{1}{2}\delta)^2 \|\mu_i\|\|\mu_j\|} \\
&\geq \frac{\mathbb{E}((z_{y_i}^m)^\top z_{y_j}^m) + \mathbb{E}((z_{y_i}^m)^\top \boldsymbol{\varepsilon}) + \mathbb{E}((z_{y_j}^m)^\top \boldsymbol{\varepsilon}) + \mathbb{E}(\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon})}{\|\mu_i\|\|\mu_j\|} \\
&\geq \frac{\mathbb{E}(\langle z_{y_i}^m, z_{y_j}^m\rangle) + \mathbb{E}(\langle z_{y_i}^m, \boldsymbol{\varepsilon}\rangle) + \mathbb{E}(\langle z_{y_j}^m, \boldsymbol{\varepsilon}\rangle) + \mathbb{E}(\langle \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}\rangle)}{\|\mu_i\|\|\mu_j\|} \\
&\geq \frac{\mu_i \cdot \mu_j + 0 + 0 + \sigma^2}{\|\mu_i\|\|\mu_j\|} = \frac{\mu_i \cdot \mu_j + \sigma^2}{\|\mu_i\|\|\mu_j\|} \\
&\geq \frac{\mu_i \cdot \mu_j}{\|\mu_i\|\|\mu_j\|} = \mathbb{E}(\langle z_{y_i}^m, z_{y_j}^m\rangle)
\end{aligned}
\tag{21}
$$

This completes the proof. □

A.6 Training Pipeline

In this subsection, we depict the training pipeline in Algorithm 1.

A.7 Additional Results.

Hyperparameter results with respect to l. As shown in Fig. 10, we analyze the sensitivity of ManiGest to the choice of the reference layer l for intra-modal geometry preservation. Across different corruption types, including *Gauss.*, *Impul.*, and *Snow*, the best performance is consistently achieved when using the shallowest layer, i.e., $l = 1$, as the geometric reference. As l increases, the accuracy generally decreases, indicating that deeper layers are less suitable as stable references for preserving pretraining-induced intra-modal geometry. This observation is consistent with prior findings that shallow layers tend to retain more general and transferable pretrained knowledge, whereas deeper layers are more task-specific and more easily affected during adaptation. Therefore, using shallow-layer geometry as the reference provides a more reliable constraint for preserving intra-modal structures under the modal-specific distribution shift.

Complete Component Ablation Results. To complement the summarized ablation study in the main text, we report the full results across all video corruption types in Table 5. The detailed results confirm that BGSR and SIGP provide consistent improvements across different corruption settings. When both components are applied, ManiGest achieves the highest average accuracy with both READ and ABPEM, further supporting the effectiveness of the proposed design.

Hyperparameter Sensitivity on Kinetics50-C. To provide a more comprehensive view of the hyperparameter sensitivity, we present the complete grid-search results of READ+ManiGest on Kinetics50-C in Fig. 11 and Fig. 12. Specifically, we vary the balancing coefficients α and β and report the corresponding classification accuracy under all audio and video corruption types. These results complement the representative examples shown in the main text.

Algorithm 1 Optimization Procedure of ManiGest

Require: Minibatch $\{\boldsymbol{x}_n = (x_n^a, x_n^v)\}_{n=1}^{N^*}$; modality-specific encoders f^a, f^v; baseline MTTA objective $\mathcal{L}$; hyperparameters α, β.

Ensure: Updated model parameters.

1: Extract audio and video token representations:

$$\boldsymbol{z}_n^m = f^m(x_n^m), \quad m \in \{a, v\}.$$

2: Estimate entropy-based token weights $\omega_1(\boldsymbol{z}_{n,t}^m)$ and $\omega_2(\boldsymbol{z}_{n,t}^m)$ for each modality.

3: Construct empirical token distributions $\mathcal{P}^1_{\boldsymbol{z}_n^m}$ and $\mathcal{P}^2_{\boldsymbol{z}_n^m}$ using ω_1 and ω_2, respectively.

4: Compute the unilateral gradient-truncated Wasserstein distances:

$$\mathcal{W}\left(\mathcal{P}^1_{\boldsymbol{z}_n^a}, \mathrm{sg}(\mathcal{P}^2_{\boldsymbol{z}_n^v})\right), \quad \mathcal{W}\left(\mathcal{P}^1_{\boldsymbol{z}_n^v}, \mathrm{sg}(\mathcal{P}^2_{\boldsymbol{z}_n^a})\right).$$

5: Compute the Birkhoff geometric symmetry restoration loss:

$$\mathcal{L}_{\mathrm{BGSR}} = \frac{1}{N^*} \sum_{n=1}^{N^*} \left[\mathcal{W}\left(\mathcal{P}^1_{\boldsymbol{z}_n^a}, \mathrm{sg}(\mathcal{P}^2_{\boldsymbol{z}_n^v})\right) + \mathcal{W}\left(\mathcal{P}^1_{\boldsymbol{z}_n^v}, \mathrm{sg}(\mathcal{P}^2_{\boldsymbol{z}_n^a})\right)\right].$$

6: Construct SPD matrices $\mathbf{P}_l^m$ and $\mathbf{P}_L^m$ from shallow-layer and last-layer representations for each modality.

7: Compute the SPD-guided intra-modal geometry preservation loss:

$$\mathcal{L}_{\mathrm{SIGP}} = \sum_{m \in \{a,v\}} \mathrm{AIRM}\left(\mathbf{P}_l^m, \mathbf{P}_L^m\right).$$

8: Combine ManiGest with the baseline objective:

$$\mathcal{L}_{ManiGest} = \mathcal{L} + \alpha \mathcal{L}_{\mathrm{BGSR}} + \beta \mathcal{L}_{\mathrm{SIGP}}.$$

9: Update model parameters by minimizing $\mathcal{L}_{ManiGest}$.

Table 5: Complete component ablation results on Kinetics50-C under video corruptions. BGSR and SIGP denote the two components of ManiGest.

Method	BGSR	SIGP	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	Avg.
READ			49.4	49.7	49.0	68.0	65.1	71.2	69.0	64.5	64.4	57.4	75.5	53.6	68.3	68.0	65.1	62.5
+ManiGest	✓		51.7	51.8	51.5	69.8	67.3	71.6	70.0	66.1	67.5	64.4	76.3	56.4	70.6	69.8	67.2	64.8
+ManiGest		✓	51.5	51.1	51.3	69.2	67.3	71.5	70.1	65.9	67.5	64.2	75.8	55.3	70.6	69.6	67.3	64.5
+ManiGest	✓	✓	**51.9**	**52.0**	**51.7**	**70.0**	**67.4**	**71.7**	**70.2**	**66.3**	**67.7**	**64.5**	**76.4**	**56.6**	**70.6**	**70.0**	**67.6**	**65.0**
ABPEM			50.3	50.9	50.6	69.5	69.2	72.3	71.7	66.0	67.1	65.7	75.4	56.2	71.6	71.2	68.2	65.0
+ManiGest	✓		53.3	55.4	54.4	70.0	69.5	73.5	70.4	65.7	67.7	65.9	76.6	58.1	71.3	71.6	69.2	66.2
+ManiGest		✓	53.3	55.4	54.4	70.0	69.5	73.5	71.7	66.3	67.7	65.9	76.6	58.1	71.6	71.6	69.2	66.3
+ManiGest	✓	✓	**53.4**	**55.9**	**54.8**	**70.4**	**69.8**	**73.6**	**70.7**	**66.2**	**68.0**	**66.2**	**76.9**	**58.2**	**71.5**	**72.0**	**69.7**	**66.5**

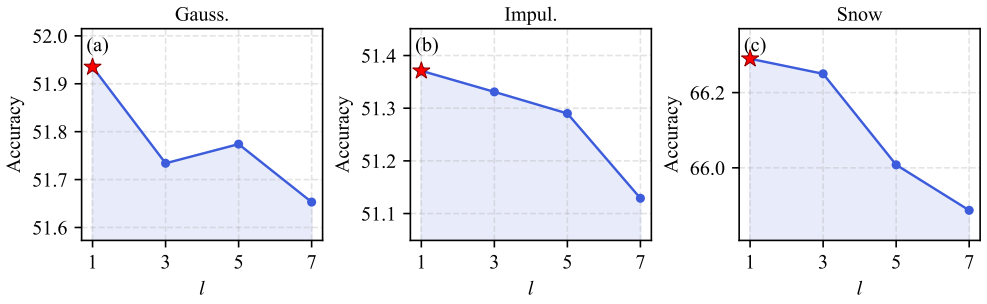

Figure 10: Effect of the layer number l on representative corruption settings in Kinetics50-C using READ as the base adaptation method. The red star indicates the best result in each subplot.

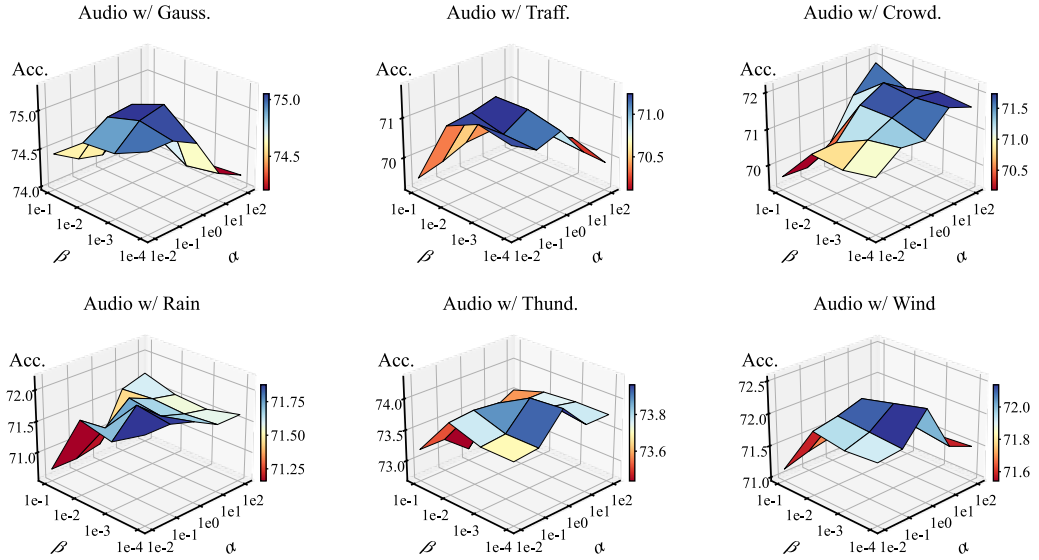

Figure 11: Complete parameter sensitivity results of READ+ManiGest on Kinetics50-C under audio corruptions. We perform a grid search over the balancing coefficients α and β, and report the classification accuracy for each audio corruption type.

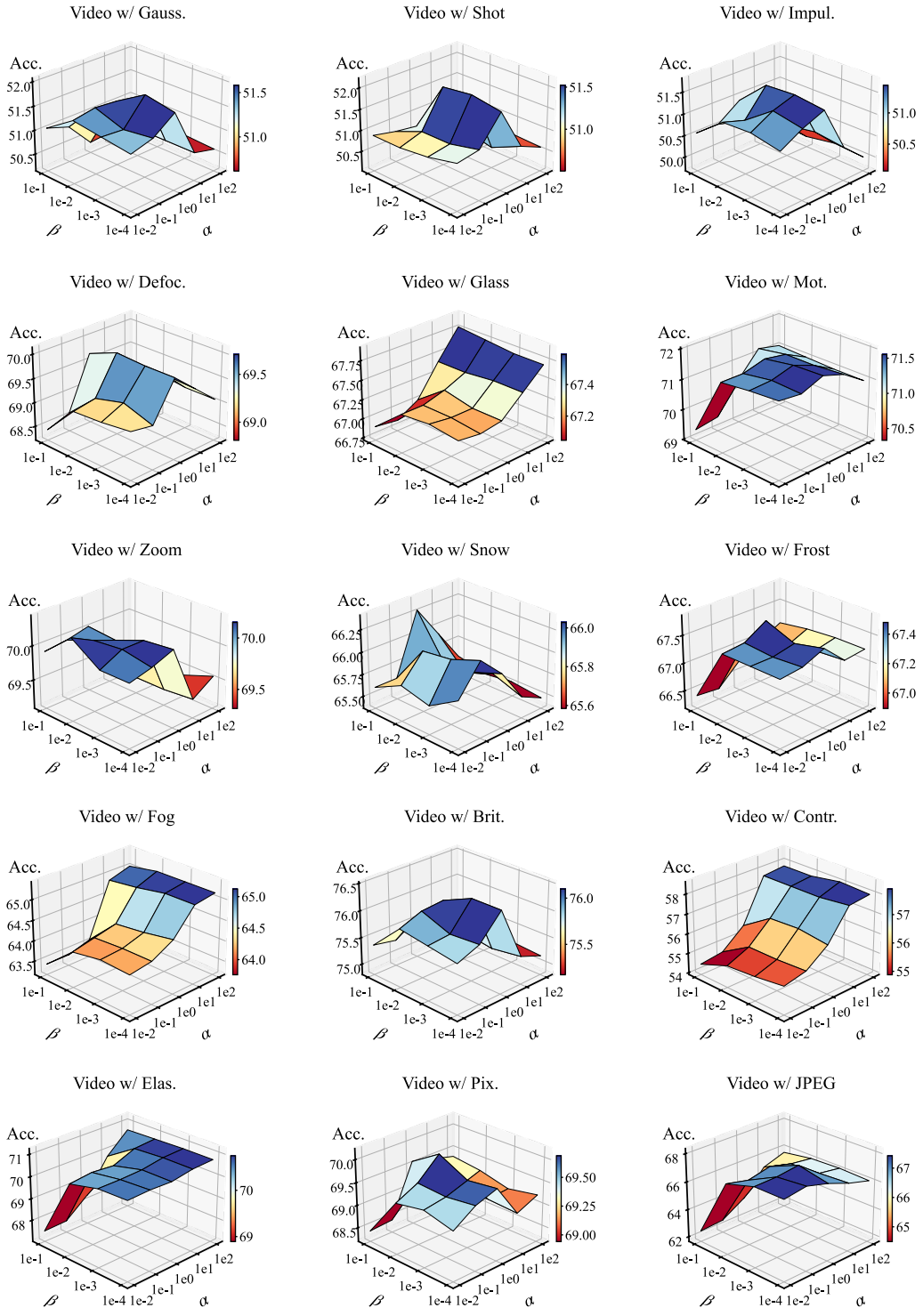

Figure 12: Complete parameter sensitivity results of READ+ManiGest on Kinetics50-C under video corruptions. We perform a grid search over the balancing coefficients α and β, and report the classification accuracy for each video corruption type.