MANIFOLD-CONSTRAINED DISTILLATION FOR MISSING-MODALITY BRAIN TUMOR SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Magnetic Resonance Imaging (MRI)-based brain tumor segmentation is crucial for clinical diagnosis and treatment planning, yet its accuracy degrades severely when part of the imaging modalities is missing. Knowledge distillation (KD) alleviates this by using a full-modality teacher to guide a modality-constrained student. Existing methods mainly rely on feature-level distillation, which incurs additional memory and computational costs, especially for high-resolution medical images. In contrast, logit-only distillation provides a lightweight alternative without explicit feature alignment, yet remains unexplored in missing-modality brain tumor segmentation. However, native logit distillation may force the student toward teacher predictions outside its reachable prediction space, as the teacher exploits modality information unavailable to the student. In this paper, we reformulate logit distillation under missing modalities as a manifold-constrained target estimation problem, and propose Manifold-Constrained Distillation (MCD), a novel geometry-aware logit distillation framework. MCD projects the teacher's correction onto the student's transferable space under the Fisher geometry, retaining only the realizable component as the target. On BraTS 2018 and BraTS 2024 benchmarks, MCD consistently improves segmentation accuracy over strong baselines across diverse missing-modality scenarios. To further demonstrate the generalizability of our framework, we extend MCD to an additional task, multimodal sentiment analysis, and again observe the superiority of our MCD model.

1 INTRODUCTION

Accurate brain tumor segmentation from multi-modal magnetic resonance imaging (MRI) is clinically important for treatment planning Bakas et al. (2017). Since different MRI modalities provide complementary information, incorporating all relevant modalities is essential for robust brain tumor segmentation Shaker et al. (2024). However, in real-world scenes, one or more MRI modalities are often unavailable due to various practical limitations Liu et al. (2023); Ding et al. (2021).

To address missing-modality brain tumor segmentation, several strategies have been explored. Existing methods fall into three paradigms Zhu et al. (2026); Azad et al. (2025). First, fusion-based methods Chen et al. (2019); Ding et al. (2021); Zhao et al. (2022b) exploit available modalities to mitigate missing modality issues, but their effectiveness depends on the design of fusion strategies Ma et al. (2021; 2022). Second, reconstruction-based methods Dorent et al. (2019); Yang et al. (2023) recover missing information by synthesising missing modalities, but are limited by reconstruction errors Wang et al. (2023). Recently, knowledge distillation (KD)-based methods Hinton et al. (2015) have demonstrated promising performance by transferring knowledge to compensate for missing modalities, including KD-Net Hu et al. (2020), M3AE Liu et al. (2023), LCKD Wang et al. (2023), MST-KDNet Zhu et al. (2025), and AdaMM Zhu et al. (2026).

Despite the success of KD-based methods, existing approaches mainly rely on feature-level distillation Zhu et al. (2025; 2026); Kang et al. (2026). However, feature distillation introduces computational overhead El-Assiouti et al. (2024), which is challenging for dense prediction tasks requiring high-resolution feature maps Fan et al. (2023) and limits the scalability of missing-modality learning to large-scale models Gu et al. (2024); Meng et al. (2023) (*e.g.*, large language models and diffusion models). Logit distillation provides a lightweight alternative and is widely adopted in large-scale model distillation Agarwal et al. (2024); Xue et al. (2026) (Fig . 1 (c)). Nevertheless, directly apply-

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.

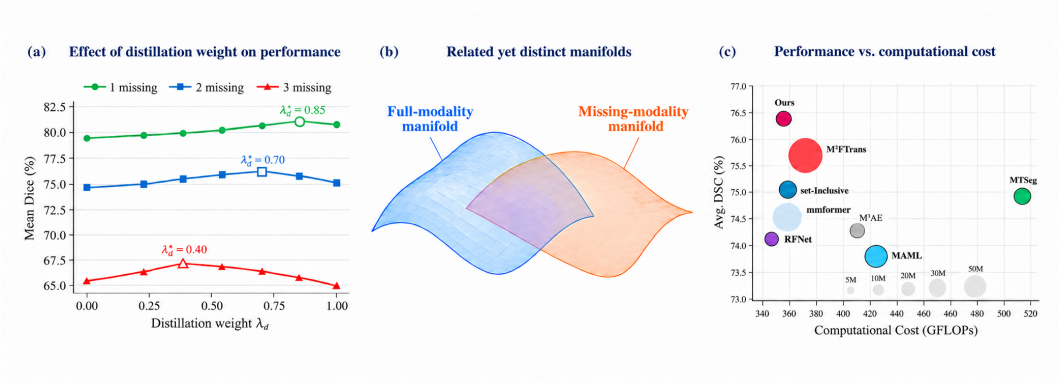


Figure 1: The motivation of the proposed correction projection. (a) Effect of the distillation weight on segmentation performance under different missing-modality settings. The optimal distillation weight decreases as severity of missing modalities increases. (b) Illustration of related yet distinct manifolds between full-modality and missing-modality inputs. (c) Comparison of segmentation performance and computational cost among different methods.

ing logit distillation to missing-modality learning remains challenging due to ineffective knowledge transfer under incomplete modalities. This raises a key question: **What limits the effectiveness of logit distillation in missing-modality scenarios?**

To answer this question, we conduct empirical analysis to investigate why conventional logit distillation is ineffective in missing-modality scenarios. Specifically, we analyze the optimal distillation weight under different missing-modality settings. As illustrated in Fig. 1 (a), the optimal KD weight decreases as the severity of missing modalities increases, indicating limited transferability of teacher knowledge under modality gaps. To explain this phenomenon, we analyze knowledge distillation from a geometric perspective based on the manifold hypothesis Fefferman et al. (2016). Full-modality and missing-modality exhibit related yet distinct manifolds (Fig. 1 (b)); therefore, missing modalities restrict the knowledge that students can infer from observable inputs Xu et al. (2022); Pournemat et al. (2021). We therefore characterize the student's reachable space using a tangent space estimated from perturbations of observable modalities Yu et al. (2019).

Building on these insights, we propose Manifold-Constrained Distillation (MCD), a geometry-aware framework for missing-modality logit distillation. MCD first estimates the student's reachable prediction space based on prediction responses induced by tangent perturbations of observable modalities. Then MCD projects teacher corrections onto this space and retains only the transferable component as the distillation target. Our contributions are summarized as follows:

- **Novel Perspective:** We provide the first systematic analysis of logit distillation, revealing that teacher supervision may fall outside the student's reachable prediction space under missing modalities from a manifold perspective.
- **Manifold-Constrained Distillation:** We propose a manifold-constrained logit distillation (MCD) framework, including a Modality-Attributed Tangent Space Estimation module to characterize the student's reachable prediction space and a Fisher-Guided Correction Projection module to retain transferable teacher knowledge while removing unreachable supervision.
- **Superior Performance and Compatibility:** Extensive experiments across various missing-modality settings demonstrate that MCD consistently achieves state-of-the-art performance while exhibiting superior computational efficiency compared with existing methods.

2 Related Work

2.1 Brain Tumor Segmentation with Missing Modalities.

Existing studies on incomplete multi-modal brain tumor segmentation mainly follow three directions Zhu et al. (2026); Azad et al. (2025): *feature fusion*, *cross-modal reconstruction*, and *knowl-*

edge distillation. Feature fusion methods exploit the available modalities through statistical feature aggregation Dorent et al. (2019), modality disentanglement Chen et al. (2019); Vadacchino et al. (2021); Ding et al. (2021), and shared-specific feature modeling Zhang et al. (2022); Qiu et al. (2023); Zhao et al. (2022b); Karimijafarbigloo et al. (2024). Cross-modal reconstruction methods recover unavailable modalities from the observed inputs using conventional image synthesis Van Tulder & de Bruijne (2015); Jog et al. (2017) or deep generative models Yu et al. (2018); Sharma & Hamarneh (2019); Dorent et al. (2019); Yang et al. (2023). Knowledge distillation methods compensate for missing information by transferring knowledge from complete-modality models to incomplete-modality models Hu et al. (2020); Wang et al. (2023); Vadacchino et al. (2021); Zhu et al. (2025) or by maximizing mutual information across modalities Liu et al. (2023). Among these paradigms, knowledge distillation offers an effective way to improve incomplete-modality models while avoiding modality reconstruction errors Hinton et al. (2015); Hu et al. (2020); Wang et al. (2023). Despite these advantages, existing methods rely on feature-level knowledge distillation, which adds computational overhead, especially for high-resolution medical images.

2.2 KNOWLEDGE DISTILLATION.

Existing Knowledge distillation methods can be broadly divided into feature-based and logit-based distillation Gou et al. (2021). Feature-based methods align representations, generally providing richer supervision but requiring additional storage and computation Wang et al. (2023); Vadacchino et al. (2021); Zhu et al. (2025). In incomplete multi-modal brain tumor segmentation, existing methods typically distill either intermediate features alone or both intermediate features and prediction logits to compensate for missing information Hu et al. (2020); Wang et al. (2023); Vadacchino et al. (2021); Zhu et al. (2025). Logit-based methods transfer knowledge using output predictions and have been further advanced by decoupling the classical distillation objective Zhao et al. (2022a); Chen et al. (2026), exploiting multi-level correlations among logits Zhao et al. (2022a), and calibrating teacher and student logits Zhu et al. (2025). Despite the progress of logit distillation, its application to missing-modality segmentation remains limited. Directly transferring teacher logits to incomplete-modality students is challenging due to the mismatch between complete and incomplete inputs. Recent on-policy distillation methods adapt teacher supervision according to student capacity Xue et al. (2026), but do not consider the modality-specific constraints caused by missing inputs. In contrast, we analyze logit distillation from a manifold perspective, estimate the student's reachable prediction space from the observable modalities, and project teacher supervision onto this space to retain transferable knowledge while removing unreachable supervision.

3 ANALYSIS AND MOTIVATION

Let $x_{\mathcal{S}}$ and $x_{\bar{\mathcal{S}}}$ denote the observed and missing modalities. The KD procedure can be formulated as follows: $z_T = f(x_{\mathcal{S}}, x_{\bar{\mathcal{S}}})$ and $z_S = P_{\mathcal{S}}(x_{\mathcal{S}}; \theta)$ are the logits of the teacher $f(\cdot)$ and the student $P_{\mathcal{S}}(\cdot; \theta)$, and $p_T^* = \sigma(z_T/\tau)$, $p_S = \sigma(z_S/\tau)$; the student minimizes $\mathrm{KL}(p_T^* \| p_S)$ as the distillation loss. Throughout this section, $\|v\|_F^2 := v^\top F v$ denotes the Fisher metric of Section 4.3, $\mu(x_{\mathcal{S}}) := \mathbb{E}[z_T \mid x_{\mathcal{S}}]$ is the teacher supervision predictable from the observed modalities, $\varepsilon := z_T - \mu(x_{\mathcal{S}})$ the remaining component unpredictable from $x_{\mathcal{S}}$, and $\Delta z := z_T - z_S$ the teacher correction.

Proposition 1. *For any student prediction z_S, $\mathbb{E}\|z_S - z_T\|_F^2 = \mathbb{E}\|z_S - \mu(x_{\mathcal{S}})\|_F^2 + \mathbb{E}\|\varepsilon\|_F^2$; the cross term vanishes as $\mathbb{E}[\varepsilon \mid x_{\mathcal{S}}] = 0$: no function of the observed modalities can predict ε, so the second term is a noise floor rather than learnable signal.*

Thus, ε is irreducible target noise. Removing the removable part in KD requires knowing which corrections the student can express: let $\mathcal{T}_\theta(x_{\mathcal{S}})$ be the locally reachable prediction space of Eq. 3 *i.e.* the first-order realizable response directions, estimated as $\widehat{\mathcal{T}}_{\mathcal{S}}$ in Section 4.2. Let Π_F denote the F-orthogonal projector onto $\mathcal{T}_\theta(x_{\mathcal{S}})$. The analysis below idealizes $\widehat{\mathcal{T}}_{\mathcal{S}} = \mathcal{T}$; the accuracy of the estimator is discussed in Appendix C. Projections act on displacements from z_S: $\Delta z = \Delta z_\parallel + \Delta z_\perp$ with $\Delta z_\parallel := \Pi_F \Delta z$, and the same convention applies to ε and $\mu - z_S$.

Proposition 2. *We write $\sigma_{\varepsilon_\parallel}^2 := \mathbb{E}\|\varepsilon_\parallel\|_F^2$, $\sigma_{\varepsilon_\perp}^2 := \mathbb{E}\|\varepsilon_\perp\|_F^2$, and $\delta^2 := \mathbb{E}\|(\mu - z_S)_\perp\|_F^2$.*

(i) Variational characterization. *The solution of* $t^* = \arg\min_t \ \mathbb{E}\|t - z_T\|_F^2 + \gamma\,\mathbb{E}\|(t - z_S)_\perp\|_F^2$ $(\gamma \geq 0)$ *is* $t^* = z_S + \Delta z_\parallel + \Delta z_\perp/(1+\gamma)$*: vanilla distillation is* $\gamma = 0$*, and the projected target is the limit* $\gamma \to \infty$*.*

(ii) Improvement and denoising. *For* $t^* = z_S + \Delta z_\parallel$*,*

$$\mathbb{E}\|t^* - \mu\|_F^2 = \mathbb{E}\|z_T - \mu\|_F^2 - \left(\sigma_{\varepsilon_\perp}^2 - \delta^2\right). \tag{1}$$

Thus t^* is strictly closer to the optimal target $\mu(x_\mathcal{S})$ whenever $\sigma_{\varepsilon_\perp}^2 > \delta^2$, and its sensitivity to the missing modalities drops by exactly $\sigma_{\varepsilon_\perp}^2$. The detailed derivation are provided in Appendix C.

Assumption 1. *In Fisher-whitened coordinates, training is locally equivalent to ordinary least squares with* n *case-level observations and* d *reachable directions under random design. The student-observable mean is well specified, target noise is conditionally zero-mean with homoscedastic Fisher variance, and projection removes* $\varepsilon_\perp$ *while incurring squared bias* δ^2*; the* $\parallel$ */* $\perp$ *split is defined in the population.*

Theorem 1. *Under Assumption 1, MCD retains* $\varepsilon_\parallel$ *(and* $\nu_\parallel$*), removes* $\Delta z_\perp$*, and, with* $\mathcal{R}(w) := \frac{1}{n}\mathbb{E}\|X(w - w^*)\|^2$ *the excess risk relative to the pattern-optimal target,*

$$\mathbb{E}_\mathcal{D}\left[\mathcal{R}(\hat{w}_{\mathrm{KD}})\right] - \mathbb{E}_\mathcal{D}\left[\mathcal{R}(\hat{w}_{\mathrm{MCD}})\right] \ \geq \ \frac{\sigma_{\varepsilon_\perp}^2\, d}{n} - \delta^2, \tag{2}$$

strictly positive whenever $\sigma_{\varepsilon_\perp}^2 > n\delta^2/d$*.*

The detailed derivation and proof are provided in Appendix C. Both conditions hold in the missing-modality regime—$\sigma_{\varepsilon_\perp}^2$ grows with the number of missing modalities while δ^2 shrinks as $\mathcal{T}_\theta(x_\mathcal{S})$ expands. These results motivate our method: the teacher's correction is projected onto the student's locally realizable directions, and $t^* = z_S + \Delta z_\parallel$ replaces z_T in the distillation objective.

4 METHOD

4.1 OVERALL FRAMEWORK

With the analysis and motivation above, this section details the proposed MCD framework, as illustrated in Fig. 2 (a): Given a full-modality teacher $f(\cdot)$ and an incomplete-modality student $P_\mathcal{S}(\cdot;\theta)$, the teacher produces logits $z_T = f(x_\mathcal{S}, x_{\bar{\mathcal{S}}})$, while the student predicts $z_S = P_\mathcal{S}(x_\mathcal{S};\theta)$ using only the observed modalities. Unlike conventional logit distillation that directly transfers the teacher prediction, we observe that the teacher correction $\Delta z = z_T - z_S$ contains both student-accessible and inaccessible components due to missing modalities. MCD therefore estimates the student observable tangent space, performs Fisher-guided correction projection.

4.2 MODALITY-ATTRIBUTED TANGENT SPACE ESTIMATION

Due to the asymmetric modality availability, the teacher provides supervision from a richer observation space than the student. We therefore characterize the locally reachable prediction variations of the student under its observable modalities, *attributed to each individual modality*, so that the characterization can be reused across missingness patterns and each retained direction can be traced back to the modality that affords it.

Locally Reachable Prediction Space. Let $J_\mathcal{S} = JP_\mathcal{S}(x_\mathcal{S};\theta)$ be the student Jacobian. The locally reachable prediction space is:

$$\mathcal{T}_\theta(x_\mathcal{S}) = \mathrm{range}\left(J_\mathcal{S}\,\big|_{\mathrm{span}\{v_{m,j}\}_{m\in\mathcal{S}}}\right), \tag{3}$$

where $\mathrm{range}(\cdot)$ denotes the set of all possible outputs. Here, we consider the prediction changes induced by label-preserving variations of the observable modalities; the corresponding tangent vectors $v_{m,j}$ are constructed below.

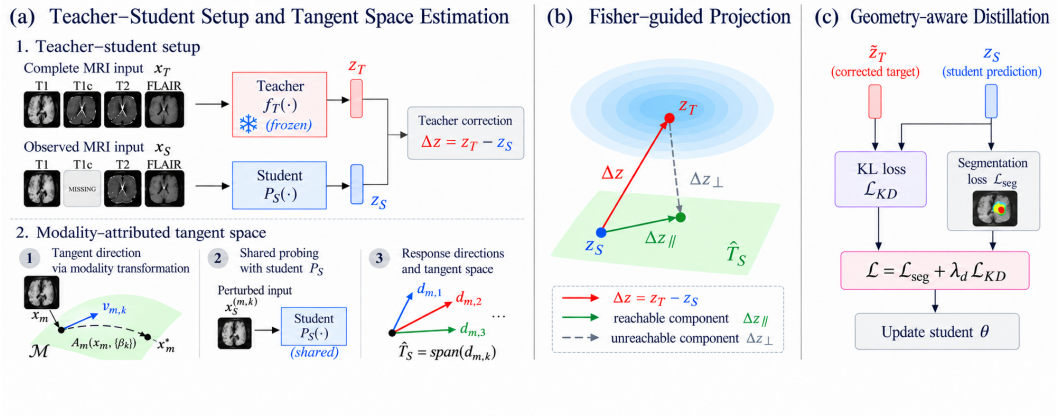


Figure 2: **Overview of MCD.** The student prediction responses induced by modality-specific tangent perturbations are collected to approximate the student-observable tangent space. MCD projects the teacher–student correction onto this space under the Fisher metric, retains the reachable component as the corrected distillation target, and optimizes the student with the geometry-aware distillation objective together with segmentation supervision.

Tangent-space Perturbation Probing. Following tangent propagation Simard et al. (2002), we take derivatives of label-preserving transformations as tangent directions and probe the student along them, one modality at a time (See Fig. 2 (a)). For each observable modality $m \in \mathcal{S}$, let $A_m(x_m, \alpha)$ be a smooth parametric transformation with $A_m(x_m, 0) = x_m$, instantiated by standard MRI appearance variations (*e.g.* low-order bias fields, B-spline elastic deformations, per-modality intensity transforms). We draw K unit vectors $\{\beta_k\}_{k=1}^K$ (random combinations of the transformation parameters, stratified across transformation types) and probe along the induced tangent directions:

$$v_{m,k} = \frac{d}{dt}\Big(x_{\mathcal{S}} + e_m \odot \big(A_m(x_m, t\beta_k) - x_m\big)\Big)\Big|_{t=0}, \qquad k = 1, \dots, K, \tag{4}$$

The student's response along each induced tangent direction is then estimated by finite differences. Specifically, we apply the transformation for each input modality at a small step $\alpha_k = \varepsilon_{\text{fd}}\beta_k$:

$$x_{\mathcal{S}}^{(m,k)} = x_{\mathcal{S}} + e_m \odot \big(A_m(x_m, \alpha_k) - x_m\big), \qquad d_{m,k} = \frac{P_{\mathcal{S}}(x_{\mathcal{S}}^{(m,k)}) - P_{\mathcal{S}}(x_{\mathcal{S}})}{\varepsilon_{\text{fd}}} \approx J_{\mathcal{S}} v_{m,k}. \tag{5}$$

Here, $x_{\mathcal{S}}^{(m,k)}$ denotes the perturbed input, and $d_{m,k}$ denotes the change in the student prediction induced by this perturbation.

Per-modality Bases and Pattern-wise Assembly. Based on Eq. 5, we can collect the responses of modality m after perturbation and obtain a set $D_m = [d_{m,1}, \dots, d_{m,K}]$. Then, we compute the modality-attributed subspace as well as the pattern-specific reachable space as:

$$\widehat{\mathcal{T}}_{\mathcal{S}} = \text{span}\Big(\bigcup_{m\in\mathcal{S}} \widehat{\mathcal{T}}^{(m)}\Big) = \text{span}\Big(\bigcup_{m\in\mathcal{S}} \text{span}(D_m)\Big), \tag{6}$$

where $\text{span}(\cdot)$ denotes the linear span operation, $\widehat{\mathcal{T}}^{(m)}$ denotes the modality-attributed subspace, and $\widehat{\mathcal{T}}_{\mathcal{S}}$ denotes the estimated reachable space.

Via Eq. 3, $\widehat{\mathcal{T}}_{\mathcal{S}}$ can approximate $\mathcal{T}_\theta$. Since $d_{m,k} \approx J_{\mathcal{S}} v_{m,k}$ for small ε_{fd} (Eq. 5), $\widehat{\mathcal{T}}^{(m)}$ captures the first-order prediction variations the student can realize through modality m, and $\widehat{\mathcal{T}}_{\mathcal{S}}$ approximates its reachable response space under pattern $\mathcal{S}$. In practice, the columns of $\bigcup_{m\in\mathcal{S}} D_m$ are orthonormalized under the Fisher metric before entering the projection of Section 4.3.

4.3 Fisher-guided Projection of Teacher Correction

The estimated tangent space characterizes the prediction variations induced by observable-modality changes. However, different prediction directions may contribute differently to the segmentation

objective. We therefore perform the projection under the Fisher metric to account for the local geometry of the student prediction (See Fig. 2 (b)).

Let $B = [b_1, \ldots, b_M]$ denote the basis of the student observable response tangent space, obtained by Fisher-orthonormalizing the stacked responses D via eigendecomposition and retaining the top M directions, such that $B^\top F B = I_M$. For each voxel i, the Fisher information metric and its global aggregation are defined as:

$$F_i = \mathrm{diag}(p_S(i)) - p_S(i)p_S(i)^\top, \qquad p_S(i) = \sigma(z_S(i)/\tau), \qquad F = \mathrm{diag}_i(F_i), \tag{7}$$

where F_i measures the local sensitivity of the segmentation distribution to prediction changes.

Given the teacher correction $\Delta z = z_T - z_S$, we seek its closest component within the student observable tangent space under the Fisher geometry:

$$\hat{c} = \arg\min_c (\Delta z - Bc)^\top F (\Delta z - Bc) = (B^\top F B)^{-1} B^\top F \Delta z, \tag{8}$$

The projected correction and its residual are then:

$$\Delta z_\parallel = B\hat{c}, \qquad \Delta z_\perp = \Delta z - \Delta z_\parallel. \tag{9}$$

Here, $\Delta z_\parallel$ represents the teacher correction expressible through student-observable prediction variations, whereas $\Delta z_\perp$ captures the remaining component outside the locally reachable student space.

4.4 Training Objective of Geometry-aware Distillation

As shown in Fig. 2 (c), the projected correction obtained in Eq. 9 can be used to construct the corrected teacher logits:

$$\hat{z}_T = z_S + \Delta z_\parallel. \tag{10}$$

Finally, the student parameters θ is optimized by combining the segmentation loss and the geometry-aware distillation loss:

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda_d \mathcal{L}_{\text{KD}}, \qquad \mathcal{L}_{\text{KD}} = \mathrm{KL}\left(\sigma(\hat{z}_T/\tau) \| \sigma(z_S/\tau)\right), \tag{11}$$

where $\mathcal{L}_{\text{seg}}$ denotes the segmentation loss with ground-truth labels, $\mathcal{L}_{\text{KD}}$ indicates the KL distance with geometry-aware constrain (Eq. 10), and λ_d balances these two terms.

5 Experiments

5.1 Datasets and Evaluation Metrics

We evaluate the model on two commonly used brain tumor segmentation benchmarks: BraTS 2018 Menze et al. (2014) and BraTS 2024 de Verdier et al. (2024). Each dataset contains four MRI modalities, including FLAIR, T1, T1ce, and T2. The ground-truth annotations include necrotic and non-enhancing tumor core, enhancing tumor, edema, and background. Following the BraTS evaluation protocol, these labels are grouped into three tumor regions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET).

BraTS 2018 contains 285 annotated cases while BraTS 2024 contains 1,350 annotated cases. Following previous studies Ding et al. (2021); Song et al. (2026), BraTS 2018 is split into 199 training cases, 29 validation cases, and 57 testing cases. BraTS 2024 is split into 945 training cases, 135 validation cases, and 270 testing cases. We use the Dice similarity coefficient (DSC) as the main metric to evaluate segmentation performance.

5.2 Implementation Details

MCD is based on the official code of RFNet Ding et al. (2021). The model is implemented with PyTorch and trained on four NVIDIA 4090 GPUs. For BraTS 2018, the batch size is set to 2 and the model is trained for 500 epochs, whereas for BraTS 2024, the batch size is set to 4 and the model is trained for 600 epochs. We optimize the model using Adam Kingma & Ba (2015) with an initial learning rate of 2×10^{-4} and adopt a cosine decay schedule Loshchilov & Hutter (2017).

Table 1: Missing modality segmentation results of different models on BraTS 2018 Menze et al. (2014) and BraTS 2024 de Verdier et al. (2024). For each scenario, from left to right, the four circles represent T2, T1ce, T1, and Flair, respectively. • represents a modality is present while ○ represents a modality is missing. The best results are highlighted in **bold**.

	Modality	●○○○	○●○○	○○●○	○○○●	●●○○	●○●○	○●●○	●○○●	○●○●	○○●●	●●●○	●●○●	●○●●	○●●●	●●●●	Avg.
		BraTS 2018															
WT	RFnet Ding et al. (2021)	84.94	76.36	**80.17**	83.89	**87.45**	88.46	87.77	80.00	85.66	86.10	88.09	90.26	88.69	86.58	89.73	85.61
	mmFormer Zhang et al. (2022)	85.50	76.90	77.10	86.20	86.00	80.20	88.50	86.90	89.20	88.00	88.80	90.00	89.20	87.00	88.76	85.88
	M2FTrans Shi et al. (2023)	84.17	73.83	75.01	86.47	84.92	78.76	**89.00**	87.44	89.04	88.36	89.29	90.04	89.51	87.71	90.19	85.58
	M3AE Liu et al. (2023)	85.40	76.80	77.10	86.30	86.10	80.50	88.50	87.00	89.20	88.20	89.00	**91.20**	89.40	87.30	85.30	85.82
	Set-Inclusive Baek et al. (2026)	85.58	76.58	77.00	86.43	86.13	80.35	88.64	87.12	89.29	88.44	88.94	90.52	89.63	87.42	90.33	86.16
	MCD (Ours)	**85.89**	**77.03**	77.28	**86.79**	86.55	**80.86**	88.97	**87.41**	**89.76**	**88.82**	**89.35**	90.78	**89.98**	**87.76**	**90.69**	**86.53**
TC	RFnet Ding et al. (2021)	68.03	81.40	65.17	66.44	82.38	71.96	82.91	73.39	83.21	71.59	82.91	73.96	83.00	83.36	83.04	76.85
	mmFormer Zhang et al. (2022)	68.50	80.30	67.00	65.50	81.50	82.00	72.50	73.00	69.50	81.30	82.85	74.00	81.40	82.10	82.30	76.25
	M2FTrans Shi et al. (2023)	68.88	82.86	66.45	66.49	83.58	70.93	84.51	73.99	85.06	72.43	**85.90**	75.05	**85.92**	85.21	**86.17**	78.22
	M3AE Liu et al. (2023)	68.54	81.91	65.68	66.95	82.90	72.48	83.43	73.91	83.73	72.11	83.43	74.48	83.52	83.88	83.60	77.37
	Set-Inclusive Baek et al. (2026)	69.11	82.48	66.25	67.52	83.46	73.04	83.99	74.47	84.29	72.67	83.99	75.04	84.08	84.44	84.12	77.93
	MCD (Ours)	**69.93**	**83.30**	**67.07**	**68.34**	**84.29**	**73.87**	**84.82**	**75.30**	**85.12**	**73.50**	84.82	**75.87**	84.91	**85.27**	84.95	**78.76**
ET	RFnet Ding et al. (2021)	41.30	71.80	36.70	37.95	78.20	75.90	43.90	43.70	43.85	77.90	78.95	49.80	71.20	74.80	75.85	60.12
	mmFormer Zhang et al. (2022)	40.81	76.66	38.62	37.27	78.10	79.02	42.86	46.38	43.76	80.51	79.30	48.27	78.98	78.11	76.55	61.68
	M2FTrans Shi et al. (2023)	43.15	76.61	37.92	**44.53**	80.12	78.96	37.05	47.60	43.81	79.71	79.83	47.53	80.96	**81.15**	**81.24**	62.67
	M3AE Liu et al. (2023)	40.71	73.52	38.85	37.87	74.77	75.44	42.90	46.00	43.65	76.64	75.61	47.58	75.40	74.78	74.03	59.85
	Set-Inclusive Baek et al. (2026)	42.17	72.86	37.48	38.81	79.54	77.16	44.88	44.69	44.84	79.21	80.27	50.66	72.32	76.01	76.45	61.16
	MCD (Ours)	**43.61**	**78.76**	**41.62**	40.57	**80.10**	**80.82**	**45.96**	**49.28**	**46.76**	**82.11**	**81.00**	**50.97**	**80.78**	80.11	79.33	**64.11**
		BraTS 2024															
WT	RFnet Ding et al. (2021)	81.65	76.66	77.19	85.71	83.77	79.45	87.36	83.35	87.65	88.10	88.21	88.02	88.69	84.22	88.66	84.58
	mmFormer Zhang et al. (2022)	81.81	77.10	77.80	85.66	84.00	79.85	87.52	83.62	87.72	88.12	88.37	88.12	88.64	84.48	88.65	84.76
	M2FTrans Shi et al. (2023)	81.94	78.36	78.85	86.90	84.33	80.60	88.09	83.92	88.23	88.79	88.99	88.51	89.11	84.75	89.14	85.37
	M3AE Liu et al. (2023)	83.28	79.63	80.19	87.44	85.25	81.24	88.91	84.49	89.07	89.43	89.52	89.30	89.66	85.26	90.02	86.18
	UniME Song et al. (2026)	85.37	**82.73**	82.75	89.00	**87.06**	84.19	90.34	**86.63**	90.44	90.79	**90.84**	90.65	91.18	**87.22**	91.14	88.02
	MCD (Ours)	**85.57**	82.68	**82.93**	**89.12**	87.00	**84.34**	**90.55**	86.59	**90.61**	**90.95**	90.81	**90.79**	**91.37**	87.17	**91.47**	**88.13**
TC	RFnet Ding et al. (2021)	61.90	74.31	60.43	57.53	77.39	76.88	65.32	65.29	65.73	78.46	79.35	67.21	79.67	78.43	80.07	71.20
	mmFormer Zhang et al. (2022)	63.08	74.14	60.55	57.46	78.12	77.08	66.50	67.66	68.16	78.42	80.25	69.60	80.47	79.86	81.59	72.20
	M2FTrans Shi et al. (2023)	62.26	75.89	63.15	60.66	79.22	79.10	67.14	66.75	67.55	80.20	81.43	69.32	81.33	80.23	81.89	73.07
	M3AE Liu et al. (2023)	68.33	76.12	66.40	63.14	81.66	80.06	69.27	70.18	71.32	80.65	83.21	71.87	83.64	82.38	84.68	75.53
	UniME Song et al. (2026)	68.65	79.68	68.75	66.13	82.74	**81.69**	72.83	73.14	73.51	83.14	83.84	75.09	**85.05**	**83.03**	85.08	77.49
	MCD (Ours)	**70.69**	**79.77**	**68.97**	**68.07**	**83.33**	80.94	**73.98**	**73.57**	**74.93**	**84.25**	**84.42**	**75.71**	84.94	82.39	**85.17**	**78.08**
ET	RFnet Ding et al. (2021)	57.23	73.53	57.04	53.70	75.06	76.25	61.37	60.60	61.18	77.14	78.59	63.73	77.67	76.20	78.75	68.54
	mmFormer Zhang et al. (2022)	58.54	73.29	56.63	53.53	77.49	76.55	61.92	63.25	62.57	77.17	78.92	65.42	79.18	78.64	80.24	69.56
	M2FTrans Shi et al. (2023)	54.58	84.04	45.79	49.81	85.03	85.51	55.34	56.60	60.18	84.87	85.56	60.38	84.90	85.64	85.78	70.93
	M3AE Liu et al. (2023)	63.55	73.98	61.92	60.08	79.04	77.53	64.91	65.46	66.89	78.30	80.81	67.34	81.03	79.82	82.12	72.19
	UniME Song et al. (2026)	66.00	**78.57**	66.04	64.40	80.99	80.34	69.58	69.43	69.73	81.75	82.41	70.89	82.30	81.58	82.81	75.12
	MCD (Ours)	**67.60**	78.36	**66.58**	**66.07**	**81.46**	**80.68**	**71.45**	**70.54**	**71.67**	**83.36**	**83.99**	**73.08**	**84.63**	**82.56**	**85.01**	**76.47**

During model training, the original MRI volumes are randomly cropped from $240 \times 240 \times 155$ to $96 \times 96 \times 96$ voxels to reduce memory consumption. Missing-modality cases are simulated by randomly removing selected modalities from the input. During inference, all possible non-empty modality combinations are evaluated. For full-resolution testing, we adopt a sliding-window inference strategy Chen et al. (2019).

5.3 Main Results

In Table 1, we present the comparison results between MCD and existing methods, including RFNet Ding et al. (2021), mmFormer Zhang et al. (2022), M2FTrans Shi et al. (2023), M3AE Liu et al. (2023), Set-Inclusive Baek et al. (2026), and UniME Song et al. (2026). Overall, MCD achieves the best averaged segmentation performance on both BraTS 2018 and BraTS 2024, demonstrating its robustness under missing-modality scenarios. Specifically, MCD consistently improves the segmentation performance of all three tumor regions, including WT, TC, and ET. Compared with the previous best method, MCD achieves improvements of 86.53 *vs*. 86.16, 78.76 *vs*. 77.93, and 64.11 *vs*. 61.16 in WT, TC, and ET, respectively, indicating that the proposed correction-aware distillation strategy can effectively transfer useful teacher knowledge for different tumor structures.

Furthermore, we analyze the results under individual missing-modality patterns. MCD maintains consistent advantages across different scenarios, with improvements observed in WT, TC, and ET segmentation (84.94 *vs*. 84.56, 76.55 *vs*. 75.73, and 58.96 *vs*. 56.16, respectively). In particular, when two or three modalities are missing, the performance degradation of existing methods becomes more severe, especially for TC and ET regions that rely more on complementary modality information. In contrast, MCD achieves more stable results in these challenging settings, demonstrating that the proposed method can effectively identify and transfer student-accessible corrections while reducing the influence of unavailable modality information.

Table 2: Ablation studies on probing directions K in Eq. 4 and projection design in Eq. 7. All experiments are conducted on BraTS 2018 Menze et al. (2014) with the RFNet backbone, and the Dice score is averaged over all 15 missing-modality patterns.

(a) Number of probing directions K.

K	WT	TC	ET	Avg.
2	86.12	77.91	63.31	75.78
4	**86.53**	**78.76**	**64.11**	**76.46**
6	86.42	78.18	63.87	76.16
8	86.31	78.09	63.42	75.94

(b) Projection and probe design.

Setting	WT	TC	ET	Avg.
Euclidean+Random	85.87	77.51	61.92	75.10
Euclidean+MA	86.18	78.58	63.75	76.17
Fisher+Random	85.92	77.63	62.08	75.21
Fisher+MA	**86.53**	**78.76**	**64.11**	**76.46**

It is also worth noting that MCD outperforms the latest UniME Song et al. (2026) on the BraTS 2024 dataset de Verdier et al. (2024), achieving 88.13 *vs*. 88.02, 78.08 *vs*. 77.49, and 76.47 *vs*. 75.12 Dice scores for WT, TC, and ET, respectively. This result further verifies the effectiveness of MCD in extracting transferable knowledge from the full-modality teacher.

5.4 Ablation Study

In this section, we perform several ablation studies to characterize the proposed method.

Number of Probing Directions K. We first investigate the impact of the number of probing directions K (Eq. 4) on the overall performance of MCD. All experiments are conducted on BraTS 2018 with the RFNet backbone, and the Dice score is averaged over all 15 missing-modality patterns. As shown in Table 2a, MCD achieves the best performance when $K{=}4$, reaching an average Dice of 76.46. Using fewer probes ($K{=}2$) underestimates the student-observable tangent space and drops the average Dice to 75.78, while further increasing K to 6 slightly degrades the performance to 76.16, as redundant probing directions introduce additional noise into the estimation of transferable corrections. We therefore set $K{=}4$ in all subsequent experiments.

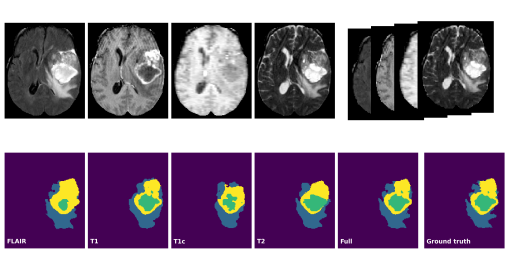
Projection Geometry and Probe Construction. Replacing the Fisher metric (Eq. 7) with Euclidean projection reduces the average Dice score from 76.46 to 76.17, further demonstrating the importance of geometry-consistent projection under the KL objective. Meanwhile, replacing the proposed modality-aware (MA) perturbation probes with random directions while keeping the Fisher metric unchanged decreases the overall performance to 75.21. This verifies that the proposed perturbations better approximate the student-observable tangent space naturally induced by available modalities. Therefore, MCD effectively combines modality-aware tangent probing with Fisher-guided projection to retain only transferable teacher corrections.

Qualitative Comparison. Fig. 3 presents qualitative results under extreme conditions, where only a single modality is provided as input. These visualizations demonstrate that our model can still produce accurate and reliable segmentations, highlighting its robustness to missing modalities. In particular, results with only the T1c modality show clearly superior delineation of the enhancing tumor region, reaffirming the importance of T1c in capturing tumor enhancement.

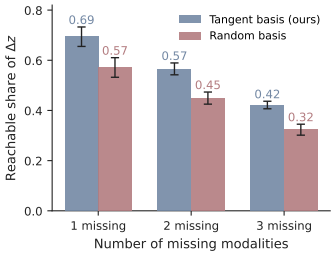
Analysis of Reachable-Space Coverage. To validate the effectiveness of our tangent perturbation, we measure the fraction of teacher correction outside the student's reachable space, $\|\Delta z_\perp\|_F^2/\|\Delta z\|_F^2$ in Eq. 9, under different missing-modality settings. As shown in Fig. 3, this fraction increases from 69% to 42% as more modalities are missing, indicating that more teacher correction becomes inaccessible to the student. Moreover, the modality-attributed tangent basis captures a larger correction fraction than a random basis with the same dimension (69% *vs*. 57%, 57% *vs*. 45% and 42% *vs*. 32%), demonstrating the effectiveness of our reachable-space estimation.

5.5 Generalization to Multimodal Sentiment Analysis

To verify that MCD generalizes beyond medical image segmentation, we further evaluate it on CMU-MOSEI Zadeh et al. (2018), a large-scale multimodal sentiment analysis dataset with aligned language, acoustic, and visual modalities. Following the standard setting Tsai et al. (2018), all



(a)



(b)

Figure 3: (a) Example segmentation results of our method on BraTS 2018 Menze et al. (2014) under four extreme single-modality settings and the full-modality setting. (b) Comparison of reachable-Space coverage under tangent basis and random basis.

Table 3: Missing-modality sentiment analysis results on the CMU-MOSEI dataset Zadeh et al. (2018). Each entry is reported in the format of Acc-2/F1 (%). The best results are highlighted in **bold**.

Models	Modalities							Avg.
	L	A	V	L,A	L,V	A,V	L,A,V	
MCTN	82.6/82.8	62.7/54.5	62.6/57.1	83.5/83.3	83.2/83.2	63.7/62.7	84.2/84.2	74.6/72.5
GCNet	83.0/83.2	60.2/60.3	61.9/61.6	84.3/84.4	84.3/84.4	64.1/57.2	85.2/85.1	74.7/73.7
MPLMM	84.6/84.6	61.8/49.4	62.2/53.2	85.1/85.1	85.2/85.1	63.2/51.0	85.3/85.3	75.3/70.5
UMDF	85.3/85.2	62.9/59.4	62.1/62.4	85.3/85.3	85.3/85.2	**65.4**/62.8	85.7/85.6	76.0/75.1
CorrKD	85.3/85.1	61.9/56.8	63.0/62.5	85.3/85.2	85.6/85.5	64.5/62.3	85.7/85.6	75.9/74.7
MMANet	84.6/84.6	62.9/48.5	**63.7/63.1**	84.8/84.8	85.3/85.2	64.5/60.9	85.3/85.2	75.9/73.2
MCD (Ours)	**85.8/85.5**	**63.3/59.9**	63.5/63.0	**85.9/85.4**	**85.4/85.5**	64.3/**64.1**	**85.7/85.2**	**76.3/75.5**

modalities are available during training, and models are tested on all non-empty modality combinations. The full-modality model serves as the teacher and modality-constrained models as students. We compare MCD with both *reconstruction-based* missing-modality methods: MCTN Pham et al. (2019), GCNet Lian et al. (2023), MPLMM Guo et al. (2024) and *distillation-based* methods: UMDF Li et al. (2024a), CorrKD Li et al. (2024b), MMANet Wei et al. (2023).

As shown in Table 3, MCD consistently outperforms these baselines, achieving an average Acc-2/F1 of 76.3/75.5 over all modality combinations. The improvement is most pronounced when only the less informative acoustic or visual modality is available (85.8/85.5 *vs*. 63.3/59.9 in the L and A columns), echoing our observation that benefits of the proposed correction projection increase with the modality gap widening. These results demonstrate that the manifold-constrained distillation perspective is task-agnostic and applies to diverse modality types beyond medical images.

6 Conclusion

This work presented MCD, a geometry-aware logit distillation framework for brain tumor segmentation with missing modalities. We first analyzed why vanilla logit distillation transfers poorly under missing inputs: a growing portion of the teacher correction lies outside the student's realizable prediction space and constitutes irreducible noise. Building on this analysis, MCD estimates the student-observable tangent space via modality-attributed probing and projects the teacher correction onto this space under the Fisher geometry, so that only transferable supervision is retained as the distillation target. Extensive experiments on two medical image segmentation datasets and one multimodal sentiment analysis dataset show that MCD consistently outperforms state-of-the-art missing-modality segmentation methods across diverse missing-modality scenarios, while introducing no additional inference cost.

AI USE STATEMENT

Generative AI tools were used only to improve the clarity, grammar, and readability of the manuscript. All technical content, including the motivation of MCD, method design, experimental protocol, result interpretation, and final claims, was developed, checked, and approved by the authors. The authors remain fully accountable for the accuracy and integrity of the paper.

REPRODUCIBILITY STATEMENT

We make efforts to ensure the reproducibility of our work. The proposed MCD framework, including modality-attributed tangent-space estimation and correction projection, is described in detail in Sec. 4, with the corresponding mathematical formulations and optimization objectives provided therein. Detailed implementation settings, including network architectures, training protocols, perturbation configurations, hyperparameters, and dataset preprocessing and missing-modality settings, are provided in the Appendix. Additional ablation studies and sensitivity analyses are also included to clarify the effects of the main design choices and hyperparameters. We will release the source code and configuration files upon publication to facilitate reproduction of the reported results.

ETHICS STATEMENT

This work uses publicly available benchmark datasets and does not involve the collection of new human-subject data. All data are used in accordance with their respective licenses and protocols. The proposed method is intended for research purposes only, and further clinical validation is required before real-world deployment.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Reza Azad, Mohammad Dehghanmanshadi, Nika Khosravi, Julien Cohen-Adad, and Dorit Merhof. Addressing Missing Modality Challenges in MRI Images: a Comprehensive Review. *Computational Visual Media*, 11(2):241–268, 2025.

Seunghun Baek, Jihwan Park, Jaeyoon Sim, Hoseok Lee, Seungjoo Lee, and Won Hwa Kim. Set-Inclusive Uncertainty Modeling for Robust Brain Tumor Segmentation. *arXiv preprint arXiv:2606.30374*, 2026.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features. *Scientific data*, 4(1):170117, 2017.

Cheng Chen, Qi Dou, Yueming Jin, Hao Chen, Jing Qin, and Pheng-Ann Heng. Robust Multimodal Brain Tumor Segmentation via Feature Disentanglement and Gated Fusion. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 447–456, 2019.

Shenghan Chen, Yiming Liu, Yanzhen Wang, Yujia Wang, and Xiankai Lu. Reframing Long-Tailed Learning via Loss Landscape Geometry. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22227–22237, June 2026.

Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, et al. The 2024 Brain Tumor Segmentation (Brats) Challenge: Glioma Segmentation on Post-treatment MRI. *arXiv preprint arXiv:2405.18368*, 2024.

Yuhang Ding, Xin Yu, and Yi Yang. RFNet: Region-aware Fusion Network for Incomplete Multi-modal Brain Tumor Segmentation. In *IEEE/CVF International Conference on Computer Vision*, pp. 3975–3984, 2021.

Reuben Dorent, Samuel Joutard, Marc Modat, Sebastien Ourselin, and Tom Vercauteren. Hetero-modal Variational Encoder-decoder for Joint Modality Completion and Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 74–82, 2019.

Omar S El-Assiouti, Ghada Hamed, Dina Khattab, and Hala M Ebied. HDKD: Hybrid Data-efficient Knowledge Distillation Network for Medical Image Classification. *Engineering Applications of Artificial Intelligence*, 138:109430, 2024.

Jiawei Fan, Chao Li, Xiaolong Liu, Meina Song, and Anbang Yao. Augmentation-free Dense Contrastive Knowledge Distillation for Efficient Semantic Segmentation. *Advances in Neural Information Processing Systems*, 36:51359–51370, 2023.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the Manifold Hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1726–1736, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015.

Minhao Hu, Matthis Maillard, Ya Zhang, Tommaso Ciceri, Giammarco La Barbera, Isabelle Bloch, and Pietro Gori. Knowledge Distillation from Multi-modal to Mono-modal Segmentation Networks. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12261, pp. 772–781, 2020.

Amod Jog, Aaron Carass, Snehashis Roy, Dzung L Pham, and Jerry L Prince. Random Forest Regression for Magnetic Resonance Image Synthesis. *Medical Image Analysis*, 35:475–488, 2017.

Ming Kang, Fung Fung Ting, Shier Nee Saw, Raphaël C-W Phan, Zongyuan Ge, and Chee-Ming Ting. A Multimodal Feature Distillation with Mamba-Transformer Network for Brain Tumor Segmentation with Incomplete Modalities. *IEEE Transactions on Artificial Intelligence*, 2026.

Sanaz Karimijafarbigloo, Reza Azad, Amirhossein Kazerouni, Saeed Ebadollahi, and Dorit Merhof. Mmcformer: Missing Modality Compensation Transformer for Brain Tumor Segmentation. In *Medical Imaging with Deep Learning*, pp. 1144–1162. PMLR, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, 2015.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. A Unified Self-distillation Framework for Multimodal Sentiment Analysis with Uncertain Missing Modalities. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 10074–10082, 2024a.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled Knowledge Distillation for Multimodal Sentiment Analysis with Incomplete Modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12458–12468. IEEE, 2024b.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph Completion Network for Incomplete Multimodal Learning in Conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8419–8432, 2023.

Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3AE: Multimodal Representation Learning for Brain Tumor Segmentation with Missing Modalities. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 1657–1665, 2023.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations*, 2017.

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

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are Multimodal Transformers Robust to Missing Modality? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18156–18165. IEEE, 2022.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On Distillation of Guided Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306. IEEE, 2023.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 6892–6899, 2019.

Ali Pournemat, Peyman Adibi, and Jocelyn Chanussot. Semisupervised Charting for Spectral Multimodal Manifold Learning and Alignment. *Pattern Recognition*, 111:107645, 2021.

Yansheng Qiu, Delin Chen, Hongdou Yao, Yongchao Xu, and Zheng Wang. Scratch Each Other's Back: Incomplete Multi-modal Brain Tumor Segmentation via Category Aware Group Self-support Learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 21317–21326, 2023.

Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. UNETR++: Delving into Efficient and Accurate 3D Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 43(9):3377–3390, 2024.

Anmol Sharma and Ghassan Hamarneh. Missing MRI Pulse Sequence Synthesis using Multi-modal Generative Adversarial Network. *IEEE Transactions on Medical Imaging*, 39(4):1170–1183, 2019.

Junjie Shi, Li Yu, Qimin Cheng, Xin Yang, Kwang-Ting Cheng, and Zengqiang Yan. MFTrans: Modality-masked Fusion Transformer for Incomplete Multi-Modality Brain Tumor Segmentation. *IEEE Kournal of Biomedical and Health Informatics*, 28(1):379–390, 2023.

Patrice Y Simard, Yann A LeCun, John S Denker, and Bernard Victorri. Transformation Invariance in Pattern Recognition—tangent Distance and Tangent Propagation. In *Neural Networks: Tricks of the Trade*, pp. 239–274. Springer, 2002.

Peibo Song, Xiaotian Xue, Jinshuo Zhang, Zihao Wang, Jinhua Liu, Shujun Fu, Fangxun Bao, and Si Yong Yeo. Uni-Encoder Meets Multi-Encoders: Representation Before Fusion for Brain Tumor Segmentation with Missing Modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15627–15638, 2026.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning Factorized Multimodal Representations. *arXiv preprint arXiv:1806.06176*, 2018.

Saverio Vadacchino, Raghav Mehta, Nazanin Mohammadi Sepahvand, Brennan Nichyporuk, James J Clark, and Tal Arbel. Had-net: A Hierarchical Adversarial Knowledge Distillation Network for Improved Enhanced Tumour Segmentation without Post-contrast Images. In *Medical Imaging with Deep Learning*, pp. 787–801. PMLR, 2021.

Gijs Van Tulder and Marleen de Bruijne. Why does Synthesized Data Improve Multi-sequence Classification? In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 531–538, 2015.

Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. Learnable Cross-modal Knowledge Distillation for Multi-modal Learning with Missing Modality. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 216–226, 2023.

Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20039–20049. IEEE, 2023.

Liming Xu, Xianhua Zeng, Bochuan Zheng, and Weisheng Li. Multi-manifold Deep Discriminative Cross-modal Hashing for Medical Image Retrieval. *IEEE Transactions on Image Processing*, 31: 3371–3385, 2022.

Leyan Xue, Feng Xiong, Mingjun Ma, and Changqing Zhang. Distill What the Student Can See: Fisher-Projected On-Policy Distillation for Vision-Language Models. *arXiv preprint arXiv:2608.01263*, 2026.

Heran Yang, Jian Sun, and Zongben Xu. Learning unified Hyper-network for Multi-modal MR Image Synthesis and Tumor Segmentation with Missing Modalities. *IEEE Transactions on Medical Imaging*, 42(12):3678–3689, 2023.

Bing Yu, Jingfeng Wu, Jinwen Ma, and Zhanxing Zhu. Tangent-normal Adversarial Regularization for Semi-supervised Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10668–10676. IEEE, 2019.

Biting Yu, Luping Zhou, Lei Wang, Jurgen Fripp, and Pierrick Bourgeat. 3D cGAN based Cross-modality MR Image Synthesis for Brain Tumor Segmentation. In *IEEE International Symposium on Biomedical Imaging*, pp. 626–630. IEEE, 2018.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal Language Analysis in the Wild: Cmu-mosei Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018.

Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmFormer: Multimodal Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 107–117, 2022.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled Knowledge Distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11953–11962, 2022a.

Zechen Zhao, Heran Yang, and Jian Sun. Modality-adaptive Feature Interaction for Brain Tumor Segmentation with Missing Modalities. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 183–192, 2022b.

Shenghao Zhu, Yifei Chen, Weihong Chen, Yuanhan Wang, Chang Liu, Shuo Jiang, Feiwei Qin, and Changmiao Wang. Bridging the Gap in Missing Modalities: Leveraging Knowledge Distillation and Style Matching for Brain Tumor Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 15967, pp. 95–106, 2025.

Shenghao Zhu, Yifei Chen, Weihong Chen, Shuo Jiang, Guanyu Zhou, Yuanhan Wang, Feiwei Qin, Changmiao Wang, and Qiyuan Tian. No Modality Left Behind: Adapting to Missing Modalities via Knowledge Distillation for Brain Tumor Segmentation. *Medical Image Analysis*, pp. 104108, 2026.

Supplementary Material

A Implementation Details

Training Settings. All experiments are implemented in PyTorch and conducted on four NVIDIA RTX 4090 GPUs. For BraTS 2018, we use a batch size of 2 and train the model for 500 epochs, while for BraTS 2024, the batch size is set to 4 and the model is trained for 600 epochs. Adam is adopted as the optimizer with an initial learning rate of 2×10^{-4} and a cosine learning rate decay schedule. During training, the MRI volumes are randomly cropped from $240 \times 240 \times 155$ to $96 \times 96 \times 96$ voxels. All non-empty modality combinations are considered during evaluation, and sliding-window inference is used for full-resolution testing.

MCD Hyperparameters. Unless otherwise specified, we use $K = 4$ probing directions for each observable modality. The tangent perturbations are constructed using low-order bias fields, B-spline elastic deformations, and per-modality intensity transformations. The finite-difference step size is set to $\epsilon_{\mathrm{fd}} = 0.05$, the distillation temperature is set to $\tau = 4$.

Perturbation Settings. For tangent-space probing, perturbations are applied to one observable modality at a time. We consider three families of label-preserving MRI variations, with the probing directions stratified across the three families. Specifically, (i) *multiplicative bias fields* are generated using low-order polynomial fields $b(\mathbf{u}) = \sum_{j=1}^{9} c_j \phi_j(\mathbf{u})$ over normalized spatial coordinates $\mathbf{u} \in [-1, 1]^3$, where $\mathbf{c} \sim \mathcal{N}(0, I_9)$; (ii) *elastic deformations* are generated from a coarse $4 \times 4 \times 4$ control grid with i.i.d. Gaussian displacements, which are trilinearly upsampled to the image resolution with border padding; and (iii) *intensity remaps* use polynomial transformations $x \mapsto x + ax^2 + bx^3$, with $(a, b) \sim \mathcal{N}(0, I_2)$. For each probe, the sampled perturbation parameters are normalized before application: unit coefficient norm is used for the bias-field and intensity transformations, while the displacement grid is normalized to unit standard deviation. The normalized perturbation is then applied with finite-difference step size $\varepsilon_{\mathrm{fd}} = 0.05$ to estimate the prediction response in Eq. equation 5. These perturbations preserve the segmentation semantics by introducing only mild appearance changes or sub-voxel spatial deformations.

B Additional Experiments

B.1 Projection versus Shrinkage

We further investigate whether the gain of MCD comes from removing unreachable correction components rather than simply reducing the magnitude of teacher supervision. We compare the full teacher correction $z_S + \Delta z$, a globally shrunk correction $z_S + \alpha \Delta z$ with α selected on the validation set, a norm-matched correction $z_S + \alpha_{\mathrm{nm}} \Delta z$ whose Fisher norm is matched to that of $\Delta z_{\|}$, and the MCD target $z_S + \Delta z_{\|}$. Specifically, $\alpha_{\mathrm{nm}} = \|\Delta z_{\|}\|_F / \|\Delta z\|_F$, so the norm-matched baseline has the same correction magnitude as MCD while retaining the original teacher direction. All variants use the same teacher, student, training schedule, and distillation objective.

B.2 Finite-Difference Step Size

We evaluate the sensitivity of MCD to the finite-difference step $\varepsilon_{\mathrm{fd}}$ used for tangent probing. Starting from the default value, we vary the step size by factors of $\{0.025, 0.05, 0.1, 0.5\}$ while keeping the perturbation directions, tangent-space construction, Fisher projection, and all other training settings fixed. This experiment tests the robustness of the local tangent approximation: an excessively small step may amplify numerical noise, whereas an excessively large step may move the perturbation outside the local linear regime.

Table 4: Projection versus correction shrinkage on BraTS 2018. Dice (%) is averaged over all 15 missing-modality patterns and reported as mean ± std over three training seeds.

Target	WT	TC	ET	Avg.
Full correction	85.75	76.92	60.91	74.53
Global shrinkage	85.94	77.42	61.19	74.85
MCD projection	86.53	78.76	64.11	76.46

Table 5: Sensitivity to the finite-difference step size on BraTS 2018. Dice (%) is averaged over all 15 missing-modality patterns.

Step size	WT	TC	ET	Avg.
0.025	86.24	78.37	63.87	76.16
0.05	86.53	78.76	64.11	76.46
0.1	86.05	78.26	64.07	76.12
0.2	85.78	77.95	63.74	75.82

C Proofs and Technical Details

C.1 Projection Identities

We use the notation and definitions introduced in Section 3. All norms and orthogonality below are with respect to the Fisher inner product

$$\langle u, v\rangle_F := u^\top F v.$$

This is the local geometry of the distillation objective, since

$$\mathrm{KL}\big(p(z+\delta z)\,\|\,p(z)\big) = \tfrac{1}{2}\,\delta z^\top F\,\delta z + O(\|\delta z\|^3), \qquad F = \mathrm{diag}(p) - pp^\top,$$

up to the global factor $1/\tau^2$ at temperature τ. If a per-voxel block is singular because of the softmax shift direction, we equivalently center logits or replace F by $F + \epsilon I$.

Let B be an F-orthonormal basis of $\mathcal{T}$, i.e. $B^\top F B = I_d$. The F-orthogonal projector onto $\mathcal{T}$ is

$$\Pi_F := B(B^\top F B)^{-1} B^\top F = B B^\top F, \qquad \Pi_F^2 = \Pi_F.$$

Consequently, any displacement v from the current prediction satisfies

$$v = v_\parallel + v_\perp, \qquad v_\parallel := \Pi_F v, \qquad v_\perp := (I - \Pi_F)v,$$

and

$$\|v\|_F^2 = \|v_\parallel\|_F^2 + \|v_\perp\|_F^2. \tag{F-Pythagoras}$$

Because z_S, F, and Π_F are deterministic given $(x_\mathcal{S}, \theta)$, linearity gives

$$\mathbb{E}[\varepsilon_\parallel \mid x_\mathcal{S}] = \Pi_F\,\mathbb{E}[\varepsilon \mid x_\mathcal{S}] = 0, \qquad \mathbb{E}[\varepsilon_\perp \mid x_\mathcal{S}] = 0. \tag{$*$}$$

C.2 Existence of Non-Transferable Supervision

Proposition 3. *If two cases share the same observed modalities but have different teacher outputs, then no function of the observed modalities can realize both teacher outputs.*

Proof. Any function $h(x_\mathcal{S})$ assigns one value to the shared observed input, whereas the teacher assigns two different values. Hence h differs from the teacher on at least one of the two cases. □

C.3 Proof of Proposition 1

By the definition $\varepsilon = z_T - \mu(x_\mathcal{S})$,

$$\begin{aligned}\mathbb{E}\|z_S - z_T\|_F^2 &= \mathbb{E}\|z_S - \mu(x_\mathcal{S}) - \varepsilon\|_F^2 \\ &= \mathbb{E}\|z_S - \mu(x_\mathcal{S})\|_F^2 + \mathbb{E}\|\varepsilon\|_F^2 - 2\,\mathbb{E}\langle z_S - \mu(x_\mathcal{S}), \varepsilon\rangle_F.\end{aligned}$$

For the cross term, $z_S - \mu(x_\mathcal{S})$ and F are deterministic given $x_\mathcal{S}$, so

$$\mathbb{E}\langle z_S - \mu(x_\mathcal{S}), \varepsilon\rangle_F = \mathbb{E}\left[\langle z_S - \mu(x_\mathcal{S}), \mathbb{E}[\varepsilon \mid x_\mathcal{S}]\rangle_F\right] = 0,$$

which proves Proposition 1.

C.4 Proof of Proposition 2(i)

Write $t = z_S + u$ and decompose $u = u_\parallel + u_\perp$. Using (F-Pythagoras), the objective becomes

$$\begin{aligned}&\mathbb{E}\|t - z_T\|_F^2 + \gamma\,\mathbb{E}\|(t - z_S)_\perp\|_F^2 \\ &\quad = \mathbb{E}\|u_\parallel - \Delta z_\parallel\|_F^2 + \mathbb{E}\|u_\perp - \Delta z_\perp\|_F^2 + \gamma\,\mathbb{E}\|u_\perp\|_F^2.\end{aligned}$$

The first term is minimized by $u_\parallel^* = \Delta z_\parallel$. For the remaining terms, differentiating with respect to $u_\perp$ gives

$$(1+\gamma)u_\perp^* = \Delta z_\perp, \qquad u_\perp^* = \frac{\Delta z_\perp}{1+\gamma}.$$

Therefore

$$t^* = z_S + \Delta z_\parallel + \frac{\Delta z_\perp}{1+\gamma}.$$

The values $\gamma = 0$ and $\gamma \to \infty$ give vanilla distillation and the projected target, respectively.

C.5 Proof of Proposition 2(ii)

Let $t^* = z_S + \Delta z_\parallel$. Since $\Delta z = (\mu - z_S) + \varepsilon$ and Π_F is linear,

$$\Delta z_\parallel = (\mu - z_S)_\parallel + \varepsilon_\parallel.$$

Hence

$$t^* - \mu = (z_S - \mu) + \Delta z_\parallel = -(\mu - z_S)_\perp + \varepsilon_\parallel.$$

The two terms are pointwise F-orthogonal, because $(\mu - z_S)_\perp \in \mathcal{T}^{\perp_F}$ and $\varepsilon_\parallel \in \mathcal{T}$. Applying (F-Pythagoras) pointwise and then taking expectations gives

$$\mathbb{E}\|t^* - \mu\|_F^2 = \delta^2 + \sigma_{\varepsilon_\parallel}^2.$$

Similarly,

$$\mathbb{E}\|z_T - \mu\|_F^2 = \mathbb{E}\|\varepsilon\|_F^2 = \sigma_{\varepsilon_\parallel}^2 + \sigma_{\varepsilon_\perp}^2.$$

Subtracting the two identities yields

$$\mathbb{E}\|t^* - \mu\|_F^2 = \mathbb{E}\|z_T - \mu\|_F^2 - \left(\sigma_{\varepsilon_\perp}^2 - \delta^2\right).$$

This proof uses only linearity and pointwise F-orthogonality; no distributional assumption on ε is needed.

Interpretation. If the teacher contains an additive privileged component $W_p u_p$ driven by factors independent of $x_\mathcal{S}$, then $\varepsilon = W_p u_p$. Thus $\varepsilon_\perp$ is the part of that missing-modality-driven supervision lying outside the student-observable space, which is the interpretation used in the main text.

C.6 Parameter-Driven versus Input-Driven Reachability

The tangent space $\mathcal{T}$ denotes input-driven reachability: output changes induced by varying the observed modalities. Parameter-driven reachability is much larger for an overparameterized student. For example, for a per-voxel head $z_i = Wh_i$, a parameter perturbation gives $\delta z_i = (\delta W)h_i$, which spans $\mathbb{R}^C$ whenever $h_i \neq 0$; deeper layers only enlarge this set. Thus the student can express directions outside $\mathcal{T}$. Those directions, however, are not predictable from $x_\mathcal{S}$, and by $(*)$ their noise component has zero conditional mean. Projection therefore restricts supervision by observability, not by model capacity.

C.7 Proof of Theorem 1

In Fisher-whitened coordinates $\tilde{z} = F^{1/2}z$, the Fisher norm becomes Euclidean and Π_F becomes an ordinary orthogonal projector. Under the local linear model in Assumption 1, the OLS estimator satisfies

$$\hat{w} - w^* = (X^\top X)^{-1} X^\top \xi.$$

Therefore, with $P_X = X(X^\top X)^{-1}X^\top$,

$$\mathcal{R}(\hat{w}) - \mathcal{R}(w^*) = \frac{1}{n}\,\xi^\top P_X \xi.$$

Since the $\parallel$ / $\perp$ split is defined in the population, $\varepsilon_\perp$ is not annihilated by the empirical projector: $P_X\varepsilon_\perp \neq 0$ almost surely. Conditional on X, each noise component of $\xi = \nu + \varepsilon_\parallel + \varepsilon_\perp$ is zero-mean and homoscedastic, the components are mutually uncorrelated, and $\mathrm{rank}(P_X) = d$ almost surely; hence

$$\mathbb{E}\big[\mathcal{R}(\hat{w}) - \mathcal{R}(w^*) \mid X\big] = \frac{\sigma^2}{n}\,\mathrm{tr}(P_X) = \frac{\sigma^2 d}{n}.$$

Vanilla KD has target-noise variance

$$\sigma^2_{\mathrm{KD}} = \sigma^2_\nu + \sigma^2_{\varepsilon_\parallel} + \sigma^2_{\varepsilon_\perp}.$$

MCD removes $\varepsilon_\perp$, retains $\sigma^2_\nu + \sigma^2_{\varepsilon_\parallel}$, and additionally incurs the projection bias $\delta^2 = \frac{1}{n}\|r\|^2$; the bias–noise cross term vanishes since r is deterministic given X and $\mathbb{E}[\xi \mid X] = 0$. Hence

$$\mathbb{E}\,\mathcal{R}(\hat{w}_{\mathrm{KD}}) - \mathbb{E}\,\mathcal{R}(\hat{w}_{\mathrm{MCD}}) = \frac{\sigma^2_{\varepsilon_\perp} d}{n} - \delta^2,$$

which is positive exactly when $\sigma^2_{\varepsilon_\perp} > n\delta^2/d$.

C.8 Accuracy of the Estimated Tangent Space

Let $\widehat{\Pi}_F$ be the projector onto the estimated tangent space, and let

$$\eta := \sin\Theta_F\big(\widehat{\mathcal{T}}_{\mathcal{S}}, \mathcal{T}_\theta(x_{\mathcal{S}})\big)$$

be the largest principal angle between the estimated and ideal subspaces, assuming matched rank.

Proposition 4. *The projected target $\hat{t}^* = z_S + \widehat{\Pi}_F \Delta z$ satisfies*

$$\|\widehat{\Pi}_F - \Pi_F\|_{F,\mathrm{op}} = \eta, \qquad \|\hat{t}^* - t^*\|_F \le \eta \|\Delta z\|_F.$$

Moreover,

$$\mathbb{E}\|\hat{t}^* - \mu\|_F^2 - \mathbb{E}\|t^* - \mu\|_F^2 \le 2\eta\sqrt{\mathbb{E}\|t^* - \mu\|_F^2\,\mathbb{E}\|\Delta z\|_F^2} + \eta^2 \mathbb{E}\|\Delta z\|_F^2.$$

Thus the conclusions of Proposition 2(ii) and Theorem 1 are stable up to an $O(\eta)$ estimation error.

Proof. For orthogonal projectors onto equidimensional subspaces, the operator norm of their difference equals the sine of the largest principal angle (the orthogonal-projector case of Davis–Kahan). Since $\hat{t}^* - t^* = (\widehat{\Pi}_F - \Pi_F)\Delta z$, the pointwise bound follows. Finally, expand

$$\|\hat{t}^* - \mu\|_F^2 = \|t^* - \mu\|_F^2 + 2\langle t^* - \mu, \hat{t}^* - t^*\rangle_F + \|\hat{t}^* - t^*\|_F^2$$

and apply Cauchy–Schwarz. □

The finite-difference probing error perturbs each retained direction by $O(\varepsilon_{fd}/\sigma_{\min})$, where $\sigma_{\min}$ is the smallest retained singular value; near-degenerate directions are removed by truncation. The retained spectrum and principal angles can therefore be checked empirically together with the K-ablation.