SELREF: SELECTIVE PRIVILEGED REFINEMENT FOR MULTIMODAL LEARNING WITH MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal learning has achieved remarkable success across a broad range of applications, yet its performance often deteriorates when one or more modalities are missing at inference. Fully observed multimodal data available during training can serve as privileged supervision. Existing distillation-based methods primarily transfer such knowledge by aligning predictions or representations between complete and incomplete inputs, leaving the usefulness of teacher guidance for each observed subset insufficiently explored. To tackle this limitation, we propose SelRef, a unified Selective Privileged Refinement framework that selectively exploits full-modality supervision to refine incomplete-modality predictions. SelRef first constructs a subset-conditioned prediction anchor by adaptively aggregating available modality experts learns an anchor-relative correction from their intermediate representations. During training, Privileged Task Gain and Current-State Compatibility jointly select teacher directions that improve upon the anchor and remain compatible with local descent at the refined prediction. We theoretically characterize the local descent properties of these directions and demonstrate the effectiveness of SelRef on three benchmarks spanning medical image segmentation and multimodal classification under complete and missing modality settings.

1 INTRODUCTION

Multimodal learning leverages complementary information from multiple modalities (Baltrušaitis et al., 2019) for accurate and robust prediction. Conventional multimodal models typically assume that all modalities are available during both training and inference (Wang et al., 2023a; Liu et al., 2023a). In real-world scenarios, however, modalities may become unavailable due to sensor failures, data corruption or protocol constraints (Neverova et al., 2016; Choi & Lee, 2019; Wang et al., 2023a; Liu et al., 2023a), leading to substantial performance degradation. This challenge has spurred extensive research on missing-modality recovery and robust multimodal representation learning.

Recovery-based methods reconstruct missing modalities or latent features from the available modalities, typically through variational completion (Dorent et al., 2019; Hamghalam et al., 2021), diffusion-based missing-modality synthesis (Meng et al., 2024), or feature reconstruction (Yuan et al., 2021; Liu et al., 2023a; Zeng et al., 2024). Despite promising advances, these approaches rely on the quality of recovered information and its preservation of task-relevant semantics. Another line of work improves robustness by learning representations across different missing-modality patterns. Among these, HeMIS (Havaei et al., 2016) handles variable modality combinations through statistical aggregation of modality-specific embeddings, while ShaSpec (Wang et al., 2023a) explicitly models shared and modality-specific representations via distribution alignment and domain classification. Transformer architectures, including mmFormer (Zhang et al., 2022) and IM-Fuse (Pipoli et al., 2025), strengthen interactions among the observed modalities. Beyond representation learning, expert-based methods adapt predictions to the observed modality subset. SimMLM (Li et al., 2025) aggregates available expert predictions and learns across nested subsets. Complete observations during training further enable privileged supervision through transferring full-modality predictions or representations (Wei et al., 2023; Wang et al., 2023b). Such transfer, however, does not explicitly assess whether the teacher provides useful guidance for a prediction formed from a particular observed subset. Under a differentiable convex reference loss, a full-modality teacher

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.

that improves upon the subset-conditioned anchor provides an anchor-relative descent direction for refinement. As refinement changes the prediction state, the same direction may no longer remain locally beneficial, motivating guidance selection based on both anchor-level gain and refined-state compatibility.

Building on this insight, we propose **Sel**ective Privileged **Ref**inement (SelRef) for Multimodal Learning with Missing Modalities, a unified framework that selectively exploits full-modality supervision to refine incomplete-modality predictions. SelRef first establishes a subset-conditioned prediction anchor by aggregating available modality experts and learns an anchor-relative correction from their intermediate representations. During training, a frozen full-modality teacher provides privileged guidance relative to the anchor. Privileged Task Gain identifies task units where the teacher improves upon the anchor, while Current-State Compatibility further evaluates whether the guidance remains locally beneficial at the refined prediction. The selected guidance determines the correction orientation, while its magnitude is learned through the downstream task objective. At inference, SelRef refines predictions based on the modalities available at test time.

The main contributions are summarized as follows:

- We introduce SelRef, a unified framework that learns to selectively refine predictions under missing modalities by constructing subset-conditioned anchors and leveraging observed-modality representations.
- We develop selective privileged guidance through Privileged Task Gain at the anchor and Current-State Compatibility at the refined prediction, and characterize their local descent properties under a differentiable convex reference loss.
- We evaluate SelRef on BraTS2023, UPMC Food-101, and MM-IMDb under complete and incomplete modality settings, demonstrating its effectiveness across brain tumor segmentation and bimodal classification tasks.

2 Related Work

Multimodal Learning with Missing Modalities. Existing methods mainly address missing modalities through missing-modality recovery and robust representation learning. Recovery-based approaches reconstruct missing modalities Dorent et al. (2019); Meng et al. (2024) or latent features Liu et al. (2023a) from the available inputs. Meanwhile, Representation-based approaches learn shared and modality-specific representations Wang et al. (2023a), or missing-aware fusion mechanisms Zhang et al. (2022); Pipoli et al. (2025) that remain effective under different missing-modality patterns. Recent studies further address modality imbalance during optimization Wang et al. (2024) and expert-based aggregation over observed modality combinations Xu et al. (2024); Li et al. (2025). Under the common setting where complete observations are available during training (Wei et al., 2023; Wang et al., 2023b; Li et al., 2025), the complete input can contain task evidence unavailable to an incomplete subset and can therefore serve as privileged supervision.

Privileged Information for Missing-Modality Learning Complete multimodal observations available during training constitute privileged information for a model that operates on incomplete inputs at inference (Vapnik & Izmailov, 2015; Lopez-Paz et al., 2016). Knowledge distillation commonly exploits this information asymmetry through a full-modality teacher. Its predictions and intermediate features supervise students receiving incomplete inputs (Hu et al., 2020; Wei et al., 2023; Wang et al., 2023b), while ProtoKD (Wang et al., 2023c) extends the transfer to pixel-wise knowledge for medical image segmentation. Recent studies broaden the structure of privileged knowledge. CorrKD (Li et al., 2024) transfers correlations across samples, categories, and responses, and CMAD (Zhuang et al., 2025) preserves feature similarities and higher-order relations with regularization adapted to modality difficulty. AdaMM (Zhu et al., 2026) further adjusts the transfer of structural and textural information according to the observed modality combination.

Despite these advances, existing methods largely align incomplete-modality predictions or representations with a full-modality teacher, leaving the usefulness of teacher guidance for each observed subset insufficiently explored. Teacher guidance from richer observations may not always benefit a particular subset prediction, and a direction that is initially favorable may become misaligned as refinement evolves. In contrast, SelRef selects privileged directions according to their task gain

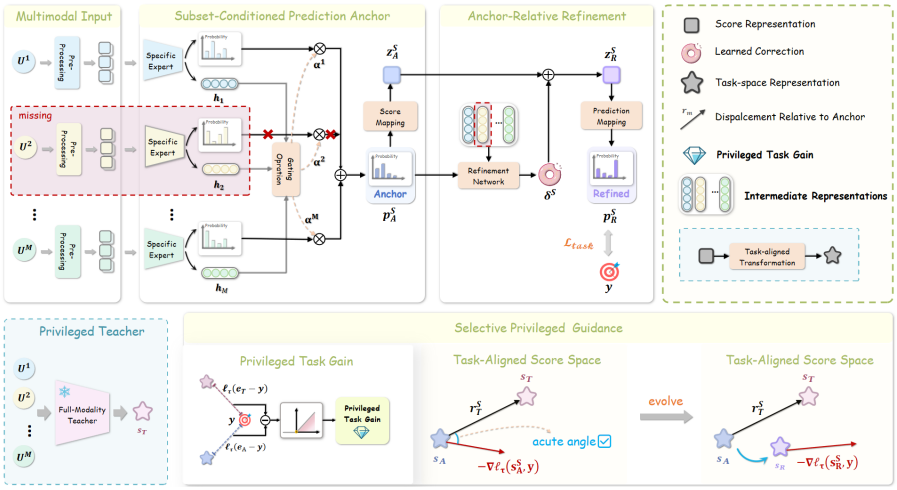

Figure 1: The overall architecture of the proposed SelRef.

over the subset-conditioned anchor and their compatibility with descent at the current prediction, ensuring that privileged supervision remains aligned with the refinement process.

3 METHODOLOGY

In this section, we introduce the SelRef framework for robust multimodal learning under missing modalities. As illustrated in Fig. 1, SelRef constructs an anchor from the available modalities and learns a subset-conditioned task correction under privileged guidance from a full-view teacher. We first formulate the problem and present the overall framework, followed by the model pipeline and optimization procedure.

3.1 PROBLEM FORMULATION

Let $\mathcal{M} = \{1, \dots, M\}$ denote the modality set. We consider a multimodal training set $\mathcal{D} = \{(\{x_i^m\}_{m\in\mathcal{M}}, y_i)\}_{i=1}^N$, where $x_i^m \in \mathcal{X}_m$ denotes the m-th modality of sample i, and $y_i \in \mathcal{Y}$ is the corresponding task target. At inference, the modalities available for sample i are indexed by $S_i \subseteq \mathcal{M}$, with $S_i \neq \varnothing$, yielding $\mathbf{x}_i^{S_i} = \{x_i^m \mid m \in S_i\}$. For simplicity, the sample index i is omitted hereafter. Given complete multimodal observations during training, we seek to learn a unified predictor $f_\theta(\mathbf{x}^S)$ that adapts to arbitrary observed modality subsets.

3.2 MODEL PIPELINE

3.2.1 SUBSET-CONDITIONED PREDICTION ANCHOR

Since different observed modality subsets provide different combinations of multimodal evidence, the predictive contribution of each modality expert varies across missing-modality scenarios. We therefore form a subset-conditioned prediction anchor that adaptively aggregates the predictions of the observed experts, providing a unified reference prediction for subsequent refinement.

For each $m \in S$, a modality-specific expert E_m produces a score $\mathbf{z}_m$ and an intermediate representation $\mathbf{h}_m$, with prediction $\mathbf{p}_m = \rho_\tau(\mathbf{z}_m)$. To contextualize expert aggregation with respect to the observed subset, we construct a compact modality descriptor $\mathbf{q}_m = \mathrm{Pool}_m(\mathrm{Proj}_m(\mathbf{h}_m))$ and summarize the current subset by $\bar{\mathbf{q}}_S = |S|^{-1}\sum_{j\in S}\mathbf{q}_j$. A subset-conditioned scoring network then computes $a_m^S = G_A(\mathbf{q}_m, \bar{\mathbf{q}}_S, \mathbf{u}_m)$, where $\mathbf{u}_m$ is a modality embedding. The scores over the observed modalities are normalized as $\boldsymbol{\alpha}^S = \mathrm{Softmax}((a_m^S)_{m\in S})$, the prediction anchor is

$$\mathbf{p}_A^S = \sum_{m\in S} \alpha_m^S \mathbf{p}_m, \qquad \mathbf{z}_A^S = \phi_\tau(\mathbf{p}_A^S), \tag{1}$$

Here, $\mathbf{p}_A^S$ denotes the subset-conditioned anchor in the prediction space, and ϕ_τ maps it back to an additive score representation $\mathbf{z}_A^S$ used by the refinement network. Each α_m^S is the normalized aggregation weight assigned to expert m under the observed subset S.

3.2.2 ANCHOR-RELATIVE REFINEMENT

The expert predictions establish compact task-level summaries, whereas their intermediate representations may retain complementary information that is not fully reflected in these outputs. This motivates an anchor-relative refinement that exploits the complementary information retained in the observed representations while preserving the subset-conditioned prediction as the reference for the correction. Since direct additive updates in the prediction space may violate task-specific output constraints, we perform the correction in an additive score space.

Given the observed representations, the anchor prediction, and the available modality subset, the refinement network R_θ predicts an anchor-relative correction, which updates the anchor as

$$\boldsymbol{\delta}^S = R_\theta\big(\{\mathbf{h}_m\}_{m\in S}, \mathbf{p}_A^S, S\big), \qquad \mathbf{z}_R^S = \mathbf{z}_A^S + \boldsymbol{\delta}^S, \qquad \mathbf{p}_R^S = \rho_\tau(\mathbf{z}_R^S). \tag{2}$$

where $\boldsymbol{\delta}^S$ represents the correction inferred from the observed modality representations relative to the anchor. Conditioning on the observed representations and modality subset enables the same refinement network to accommodate different missing-modality patterns. We initialize the output layer of R_θ to zero, such that $\boldsymbol{\delta}^S \approx \mathbf{0}$ at the start of refinement training. Accordingly, the refinement network therefore is initialized at the subset-conditioned anchor and subsequently optimized by the downstream task objective and the privileged supervision introduced next.

3.2.3 SELECTIVE PRIVILEGED GUIDANCE

Complete multimodal inputs available during training can provide task-relevant evidence beyond an observed subset S. We capture this additional information with a *Full-Modality Teacher* T, trained on complete inputs and frozen during refinement optimization, and denote its output score by $\mathbf{z}_T = T(x^{\mathcal{M}})$.

SelRef first assesses whether the full-modality teacher improves upon the subset-conditioned anchor in a task-aligned score space. When the full-modality teacher achieves a lower reference loss than the anchor, the corresponding anchor-to-teacher displacement defines a descent direction at the anchor, as shown in Proposition 1. This conclusion, however, is specific to the anchor. As refinement moves the prediction from $\mathbf{s}_A^S$ to $\mathbf{s}_R^S$, the local loss gradient may change, and following the same teacher direction may no longer reduce the reference loss locally. We therefore evaluate its directional compatibility at the current refined prediction before applying the privileged alignment objective.

Task-Aligned Score Space. The correction in Eq. 2 is additive in score space, but softmax predictions are invariant to a common shift of all class logits. Such shifts can enter a raw-score displacement without changing the prediction. We therefore evaluate privileged guidance after a fixed task-dependent mapping Γ_τ, defining $\mathbf{s}_A^S = \Gamma_\tau(\mathbf{z}_A^S)$, $\mathbf{s}_R^S = \Gamma_\tau(\mathbf{z}_R^S)$, and $\mathbf{s}_T = \Gamma_\tau(\mathbf{z}_T)$.

For mutually exclusive classification, Γ_τ centers the class logits to remove this invariant component. Independently parameterized binary outputs retain their logits. For BraTS segmentation, Γ_τ maps voxelwise four-class logits to the log-odds of the whole tumor (WT), tumor core (TC), and enhancing tumor (ET) regions, matching the region targets used for supervision. The task units $u \in \mathcal{U}_\tau$ are examples for single-label classification, example–label pairs for multi-label classification, and voxels for segmentation. A reference loss $\ell_\tau(\mathbf{s}_u, y_u)$, differentiable and convex in the task-aligned score $\mathbf{s}_u$, evaluates each unit. For BraTS, it averages binary cross-entropy over the three region scores. The loss difference between anchor and teacher defines gain, while the loss gradient at the refined prediction enters the compatibility test. More related details are defered to Appendix B.2.

Privileged Task Gain. The full-modality teacher provides additional evidence beyond the subset-conditioned anchor, but its guidance may not be equally beneficial across samples or task units. We therefore use the task-aligned reference loss to assess whether it improves upon the subset-conditioned anchor before leveraging its privileged guidance for refinement. For each task unit $u \in \mathcal{U}_\tau$, we define the *Privileged Task Gain* as

$$g_u^S = \ell_\tau(\mathbf{s}_{A,u}^S, y_u) - \ell_\tau(\mathbf{s}_{T,u}, y_u), \tag{3}$$

A positive g_u^S indicates that the full-modality teacher attains a lower reference loss than the subset-conditioned anchor for task unit u, and hence provides a favorable privileged prediction relative to the evidence available under S.

Information-Theoretic Interpretation. For a likelihood-based reference loss $\ell_\tau(\mathbf{s}, y) = -\log q_\tau(y \mid \mathbf{s})$, the expected Privileged Task Gain admits an interpretation through the mutual information between the task target and the task-aligned scores. Fix an observed subset S and task unit u, and let $\mathcal{R}_{A,u}^S$ and $\mathcal{R}_{T,u}$ denote the reference losses of the subset-conditioned anchor and the full-modality teacher averaged over the data distribution. Each expected loss defines a variational lower bound on the mutual information between Y_u and the corresponding task-aligned score: $\underline{\mathcal{I}}_{A,u}^S = H(Y_u) - \mathcal{R}_{A,u}^S$ and $\underline{\mathcal{I}}_{T,u} = H(Y_u) - \mathcal{R}_{T,u}$. Their difference is

$$\underline{\mathcal{I}}_{T,u} - \underline{\mathcal{I}}_{A,u}^S = \mathcal{R}_{A,u}^S - \mathcal{R}_{T,u} = \mathbb{E}[g_u^S]. \tag{4}$$

A positive expected gain thus yields a higher variational lower bound for the teacher prediction than for the subset-conditioned anchor. During training, guidance is selected using the observed loss difference g_u^S for each task unit. To understand the resulting selection at the individual-unit level, we next examine the local geometry of the reference loss. Let $\mathbf{r}_T^S = \mathbf{s}_T - \mathbf{s}_A^S$ denote the teacher displacement relative to the anchor. Under the convexity of the reference loss, a positive g_u^S certifies that $\mathbf{r}_{T,u}^S$ is a local descent direction at the anchor.

Proposition 1 (Local Descent Direction at the Anchor). Assume that $\ell_\tau(\mathbf{s}, y)$ is differentiable and convex with respect to $\mathbf{s}$. If $g_u^S > 0$, then

$$\nabla_{\mathbf{s}} \ell_\tau(\mathbf{s}_{A,u}^S, y_u)^\top \mathbf{r}_{T,u}^S < 0. \tag{5}$$

The proof is given in Appendix B.3. We therefore retain only positive gains and convert them into a bounded guidance weight $w_{\mathrm{gain},u}^S = \mathrm{sg}\left[\frac{[g_u^S]_+}{[g_u^S]_+ + \tau_g}\right]$, with $[a]_+ = \max(a, 0)$, where $\mathrm{sg}[\cdot]$ denotes stop-gradient and $\tau_g > 0$ controls the saturation scale.

The proposition determines the locally beneficial orientation of the full-modality correction at $\mathbf{s}_{A,u}^S$. Since the teacher and the refinement network operate with different modality information, the refinement magnitude is left to the downstream task objective rather than being inherited from the teacher displacement, while the teacher displacement provides its privileged directional guidance.

Current-State Compatibility. Proposition 1 establishes a local descent property of the positive-gain teacher displacement at the subset-conditioned anchor. The privileged guidance, however, acts on the refinement whose current prediction is $\mathbf{s}_R^S$. As refinement changes the prediction to $\mathbf{s}_R^S$, the same teacher displacement may no longer remain aligned with a descent direction of the reference loss. We therefore reassess the same anchor-relative teacher direction at the current refined prediction.

Let $\mathrm{dir}(\mathbf{r}) = \begin{cases} \mathbf{r}/\|\mathbf{r}\|_2, & \mathbf{r} \neq \mathbf{0}, \\ \mathbf{0}, & \mathbf{r} = \mathbf{0}, \end{cases}$ and define, for each task unit $u \in \mathcal{U}_\tau$,

$$\mathbf{d}_{T,u}^S = \mathrm{dir}(\mathbf{r}_{T,u}^S), \qquad c_u^S = \cos_\epsilon\left(-\nabla_{\mathbf{s}} \ell_\tau(\mathbf{s}_{R,u}^S, y_u), \mathbf{d}_{T,u}^S\right). \tag{6}$$

where $\cos_\epsilon(\cdot, \cdot)$ denotes cosine similarity with a small $\epsilon > 0$ added to the denominator for numerical stability. A positive c_u^S indicates that the teacher direction forms an acute angle with the negative reference-loss gradient at $\mathbf{s}_{R,u}^S$, and hence is aligned with the local descent direction of the reference loss at the refined prediction.

Proposition 2 (Local Descent Direction at the Refined Prediction). Assume that $\ell_\tau(\mathbf{s}, y)$ is differentiable at $\mathbf{s}_{R,u}^S$. If $c_u^S > 0$, then

$$\nabla_{\mathbf{s}} \ell_\tau(\mathbf{s}_{R,u}^S, y_u)^\top \mathbf{d}_{T,u}^S < 0, \tag{7}$$

and there exists $\eta_0 > 0$ such that, for every $0 < \eta < \eta_0$,

$$\ell_\tau\left(\mathbf{s}_{R,u}^S + \eta \mathbf{d}_{T,u}^S, y_u\right) < \ell_\tau(\mathbf{s}_{R,u}^S, y_u). \tag{8}$$

The proof is given in Appendix B.3. We accordingly retain positive compatibility through the detached weight $w_{\mathrm{cmp},u}^S = \mathrm{sg}\left[[c_u^S]_+\right]$.

The teacher direction remains defined relative to the subset-conditioned anchor, as it captures the correction indicated by the full- modality teacher with respect to the prediction supported by the observed subset. Evaluating compatibility at $\mathbf{s}_R^S$ then determines whether this anchor -relative direction remains locally descending after refinement. Consequently, Privileged Task Gain identifies teacher corrections that are beneficial relative to the observed subset, while Current-State Compatibility retains those whose direction remain locally compatible with descent at the refined prediction. The selective directional objective built from these two criteria is introduced next.

Selective Directional Alignment. The preceding criteria identify teacher directions that are beneficial relative to the subset-conditioned anchor and remain locally appropriate at the current refined prediction. We use these selected directions to supervise the anchor-relative refinement in the same task-aligned score space. Specifically, we define the refinement displacement as $\mathbf{r}_R^S = \mathbf{s}_R^S - \mathbf{s}_A^S$ and obtain refinement directions $\mathbf{d}_{R,u}^S = \mathrm{dir}(\mathbf{r}_{R,u}^S)$ using the same direction operator introduced above. The teacher direction $\mathbf{d}_{T,u}^S = \mathrm{dir}(\mathbf{r}_{T,u}^S)$ is therefore compared with the refinement direction relative to the common anchor. We measure their directional discrepancy by

$$\ell_{\mathrm{dir},u}^S = 1 - \left\langle \mathbf{d}_{R,u}^S, \mathbf{d}_{T,u}^S \right\rangle. \tag{9}$$

This objective encourages the refinement displacement to align with the teacher displacement without imposing the teacher displacement magnitude.

When either displacement is close to zero, its normalized direction is poorly determined and may yield unstable directional supervision. We therefore attenuate directional supervision in such cases using $w_{\mathrm{norm},u}^S = \mathrm{sg}\left[\frac{\|\mathbf{r}_{T,u}^S\|_2}{\|\mathbf{r}_{T,u}^S\|_2+\tau_T}\frac{\|\mathbf{r}_{R,u}^S\|_2}{\|\mathbf{r}_{R,u}^S\|_2+\tau_R}\right]$, where $\tau_T, \tau_R > 0$ set the norm scales below which directional supervision is attenuated. This factor attenuates unreliable directional supervision, whereas $w_{\mathrm{gain},u}^S$ and $w_{\mathrm{cmp},u}^S$ control the contribution of the teacher direction to privileged guidance. The privileged directional loss is then

$$\mathcal{L}_{\mathrm{priv}} = \frac{1}{|\mathcal{U}_\tau|}\sum_{u\in\mathcal{U}_\tau} w_{\mathrm{gain},u}^S w_{\mathrm{cmp},u}^S w_{\mathrm{norm},u}^S \ell_{\mathrm{dir},u}^S. \tag{10}$$

The reduction is taken over all task units. Consequently, sparse guidance is not renormalized to preserve its overall magnitude, and its contribution decreases as fewer task units receive directional supervision.

3.3 Training and Inference

The refinement network is optimized with downstream task loss and the privileged directional loss

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}}\big(\mathbf{p}_R^S, y\big) + \lambda_{\mathrm{priv}}\mathcal{L}_{\mathrm{priv}}, \tag{11}$$

where λ_{priv} controls the strength of privileged guidance. The task loss supervises the refined prediction, while $\mathcal{L}_{\mathrm{priv}}$ aligns the correction direction with selected teacher directions.

We first train the modality-specific experts, then learn the subset-conditioned anchor with the experts fixed. The full-modality teacher is trained on complete inputs and frozen before refinement training. During refinement training, we sample a nonempty subset S for each example and optimize R_θ with Eq. 11. The experts and anchor remain fixed, while the teacher processes the complete input to provide privileged guidance.

At inference, the available experts construct the anchor and provide their intermediate representations to R_θ, which produces the final prediction through Eq. 2. The teacher and target-dependent guidance are used only during training.

4 Experiments

4.1 Experimental Setup

We conduct extensive experiments covering brain tumor segmentation and multimodal classification. Specifically, we evaluate segmentation performance on BraTS2023 and conduct classification experiments on UPMC Food-101, and MM-IMDb. Due to space limitations, details of missing-modality protocols, hyper-parameters setting, and baselines are deferred to the Appendix.

	Model	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	Avg.
Enhancing Tumor (ET)	U-HVED (Dorent et al., 2019)	39.87	45.17	78.37	37.75	47.99	80.68	40.36	80.91	51.08	80.80	81.79	51.75	81.71	81.16	81.95	64.09
	mmFormer (Zhang et al., 2022)	55.93	53.08	81.70	56.47	59.79	83.22	61.01	82.94	59.14	83.04	83.47	62.11	83.15	83.35	82.90	71.42
	SFusion† (Liu et al., 2023b)	52.20	48.90	82.20	54.30	57.40	83.90	59.00	83.60	57.10	83.80	83.80	60.10	83.90	84.00	84.00	70.55
	ShaSpec (Wang et al., 2023a)	55.01	51.87	80.96	55.06	58.98	82.31	59.13	82.46	57.58	82.09	82.93	60.65	82.37	82.50	82.77	70.44
	IM-Fuse (Pipoli et al., 2025)	59.41	56.27	83.52	59.26	**63.75**	84.98	64.23	84.49	**63.50**	84.44	85.11	66.11	85.16	84.59	85.22	74.00
	SimMLM (Li et al., 2025)	58.79	56.87	84.96	60.69	60.71	85.19	62.03	**86.23**	60.29	85.05	85.31	63.02	**86.21**	85.17	85.30	73.72
	AdaMM (Zhu et al., 2026)	59.75	56.97	84.92	59.22	63.21	85.08	64.16	85.02	61.25	85.33	84.87	65.18	85.12	**86.26**	85.56	74.13
	SelRef (Ours)	**61.17**	**57.89**	**85.41**	**61.12**	62.75	**85.60**	**66.40**	85.77	62.02	**85.53**	**86.10**	**67.10**	86.03	85.63	**86.06**	**74.97**
Tumor Core (TC)	U-HVED (Dorent et al., 2019)	61.80	62.93	83.75	53.74	68.38	86.56	61.07	86.21	69.00	85.53	87.23	71.02	87.49	86.35	87.43	75.90
	mmFormer (Zhang et al., 2022)	77.40	74.30	88.05	76.28	79.73	89.68	79.67	89.33	78.15	89.92	90.20	80.47	90.13	90.19	90.34	84.26
	SFusion† (Liu et al., 2023b)	74.00	70.40	86.70	74.30	77.40	88.60	77.80	88.40	76.50	89.40	89.10	78.50	89.30	89.50	89.50	82.63
	ShaSpec (Wang et al., 2023a)	75.86	73.41	88.39	75.12	79.44	89.75	78.41	89.77	77.73	89.68	90.43	79.91	90.12	90.10	90.62	83.92
	IM-Fuse (Pipoli et al., 2025)	77.78	75.31	**90.49**	76.48	80.85	**91.38**	79.88	**91.25**	79.12	91.24	**91.56**	81.35	91.36	**91.52**	91.54	85.41
	SimMLM (Li et al., 2025)	**84.70**	75.56	78.34	**81.71**	**85.55**	88.96	**86.18**	80.65	**82.33**	85.48	88.99	**86.29**	89.26	85.37	89.25	84.57
	AdaMM (Zhu et al., 2026)	79.02	75.90	89.28	77.50	81.15	91.05	80.82	90.91	78.94	91.18	91.24	81.43	91.26	91.17	91.50	85.49
	SelRef (Ours)	79.97	**77.18**	90.04	78.03	80.67	90.93	81.30	90.70	78.86	**91.45**	90.87	81.69	**91.42**	91.28	**91.97**	**85.76**
Whole Tumor (WT)	U-HVED (Dorent et al., 2019)	69.42	81.46	71.60	85.44	84.76	77.34	86.14	85.29	89.18	88.21	86.11	89.42	88.60	89.74	89.81	84.17
	mmFormer (Zhang et al., 2022)	90.65	81.00	79.44	87.37	91.69	92.06	91.86	82.73	89.02	89.38	92.18	92.14	92.60	89.78	92.61	88.97
	SFusion† (Liu et al., 2023b)	89.10	78.50	77.60	87.00	90.80	91.20	91.30	81.30	88.20	88.70	91.70	91.60	92.10	88.80	92.20	88.01
	ShaSpec (Wang et al., 2023a)	91.28	80.19	80.25	87.24	92.19	92.53	92.50	83.24	88.72	88.87	92.79	92.70	93.06	89.47	93.15	89.21
	IM-Fuse (Pipoli et al., 2025)	91.78	82.90	83.67	88.41	92.41	92.69	92.59	85.36	89.86	90.15	92.81	92.80	93.09	90.45	93.14	90.14
	SimMLM (Li et al., 2025)	92.03	82.62	82.60	88.65	92.28	92.83	92.92	85.39	89.48	89.88	92.93	93.08	93.24	90.02	93.25	90.08
	AdaMM (Zhu et al., 2026)	91.34	81.97	82.73	88.30	92.00	92.63	92.34	84.57	89.35	89.91	92.88	92.53	93.03	90.18	93.03	89.79
	SelRef (Ours)	**92.19**	**84.80**	**85.58**	**89.79**	**92.49**	**92.93**	**93.01**	**86.30**	**90.20**	**90.78**	**93.24**	**93.16**	**93.54**	**91.06**	**93.53**	**90.84**

Table 1: **Dice score (%) comparison on BraTS2023** under both complete and incomplete modality settings. Filled modality markers indicate observed modalities, whereas hollow markers indicate missing modalities. Results marked by † are taken from Pipoli et al.(Pipoli et al., 2025). Bold and underlined values indicate the best and second-best results, respectively.

Dataset	Modality	mmFormer	ShaSpec	IMFuse	SimMLM	**SelRef (ours)**
Food-101 (Accuracy)	Text	84.71	84.10	85.02	**86.27**	86.15
	Image	77.02	77.13	76.89	65.70	**77.82**
	Text-Image	94.51	94.81	94.68	92.93	**94.86**
	Average	85.41	85.35	85.53	81.63	**86.28**
MM-IMDb (F1-Macro)	Text	54.08	54.25	58.36	57.01	**58.55**
	Image	46.89	50.09	49.44	49.57	**50.94**
	Text-Image	57.59	59.29	59.98	60.25	**60.42**
	Average	52.85	54.54	55.93	55.61	**56.64**

Table 2: **Performance comparison on UPMC Food-101 and MM-IMDb under both complete and incomplete modality settings.**

Datasets. We evaluate SelRef on three multimodal benchmarks spanning brain tumor segmentation and bimodal classification tasks. For brain tumor segmentation, we conduct experiments on BraTS2023 (Baid et al., 2021; 2024) under both complete and incomplete modality settings, reporting Dice scores for whole tumor (WT), tumor core (TC), and enhancing tumor (ET). For bimodal classification, we evaluate on UPMC Food-101 (Wang et al., 2015) and MM-IMDb (Arevalo et al., 2017), using accuracy and F1-Macro as evaluation metrics, respectively.

Baselines. For brain tumor segmentation, we select baselines spanning modality recovery (U-HVED (Dorent et al., 2019)), robust representation learning and fusion (mmFormer (Zhang et al., 2022), SFusion (Liu et al., 2023b), ShaSpec (Wang et al., 2023a), and IM-Fuse (Pipoli et al., 2025)), expert-based modality adaptation (SimMLM (Li et al., 2025)), and knowledge distillation (AdaMM (Zhu et al., 2026)). The classification comparisons include mmFormer, ShaSpec, IM-Fuse, and SimMLM.

4.2 Comparison Study

Segmentation Results. Table 1 reports segmentation performance on BraTS2023 under complete and incomplete modality settings. Across all 15 modality configurations, SelRef achieves the highest average Dice for ET, TC, and WT, reaching 74.97%, 85.76%, and 90.84%, respectively. These

results exceed the best competing method by 0.84, 0.27, and 0.70 percentage points across the three tumor regions. The gains are most pronounced when only one modality is available (NM$=3$), with detailed comparisons reported in Appendix Table 7. The results also reveal differences across tumor regions. SelRef ranks first for WT in every modality configuration and for ET in 10 of the 15 configurations. The gains on TC are comparatively smaller, although SelRef still achieves the highest overall average performance. The larger improvements under NM$=3$ align with SelRef's design principle of evaluating full-modality guidance relative to the prediction formed from the available modalities, where the usefulness of privileged information depends on the observed subset.

Classification Results. We further evaluate SelRef on UPMC Food-101 and MM-IMDb to assess its generalization to bimodal classification tasks. As reported in Table 2, SelRef achieves an average accuracy of 86.28% on Food-101, improving over the best competing method by 0.75 percentage points. Its image-only accuracy reaches 77.82%, exceeding the best competing result by 0.69 points, while SimMLM achieves a slightly higher text-only accuracy (86.27% versus 86.15%). On MM-IMDb, SelRef achieves the highest F1-Macro across all three modality configurations, reaching an average of 56.64% and improving over the best competing method by 0.71 points. The largest improvement is observed in the image-only setting, where SelRef achieves 50.94% compared with 50.09% for the best competing method. These demonstrate that SelRef consistently improves robustness under incomplete modality settings while maintaining competitive performance.

4.3 Ablation Study

To gain deeper insight into the effectiveness of the key designs in SelRef, we conduct ablation studies on BraTS2023. Table 3 evaluates the contribution of anchor-relative refinement and investigates different strategies for incorporating full-modality guidance.

Variant	WT	TC	ET	Avg.
Prediction Anchor	86.78	83.47	71.39	80.55
w/o Privileged Guidance	90.71	85.12	74.43	83.42
w/ Prediction KD	90.54	85.26	73.05	82.95
w/o Guidance Selection	90.68	85.26	74.26	83.40
Full SelRef	**90.84**	**85.76**	**74.97**	**83.86**

Table 3: Ablation results of the proposed SelRef on the BraTS2023 dataset.

The Importance of Anchor-Relative Refinement. The *Prediction Anchor* variant aggregates predictions from available modality experts without refinement, achieving an average Dice of 80.55%. The *w/o Privileged Guidance* variant further learns anchor-relative correction with the downstream task objective alone, increasing the average Dice to 83.42%, with WT and ET improving from 86.78% to 90.71% and from 71.39% to 74.43%, respectively. These gains demonstrate that observed-modality representations provide complementary task-relevant information beyond expert predictions for anchor-relative refinement.

The Importance of Privileged Guidance. Compared with *w/o Privileged Guidance*, *Full SelRef* further incorporates selected teacher guidance directions to orient the learned correction. The average Dice increases from 83.42% to 83.86%, including gains of 0.54 points on ET and 0.64 points on TC. These results demonstrate that full-modality supervision provides complementary guidance beyond the information learned from observed modalities and task supervision.

The Importance of Selective Guidance. The *w/ Prediction KD* variant replaces SelRef's selective privileged guidance with prediction-level distillation from the full-modality teacher, achieving 82.95% average Dice versus 83.86% for *Full SelRef*. Although it slightly improves TC over *w/o Privileged Guidance* (85.26% versus 85.12%), its ET score decreases from 74.43% to 73.05%. The *w/o Guidance Selection* variant retains directional supervision while omitting Gain- and Compatibility-based selection, reaching 83.40% average Dice. These results demonstrate the importance of evaluating teacher guidance relative to both the subset-conditioned anchor and the refined prediction state.

4.4 Qualitative analysis of anchor-relative refinement.

To further investigate the behavior of anchor-relative refinement, we visualize representative cases comparing the subset-conditioned prediction anchor and SelRef in Fig. 2. In Cases 1 and 2, SelRef

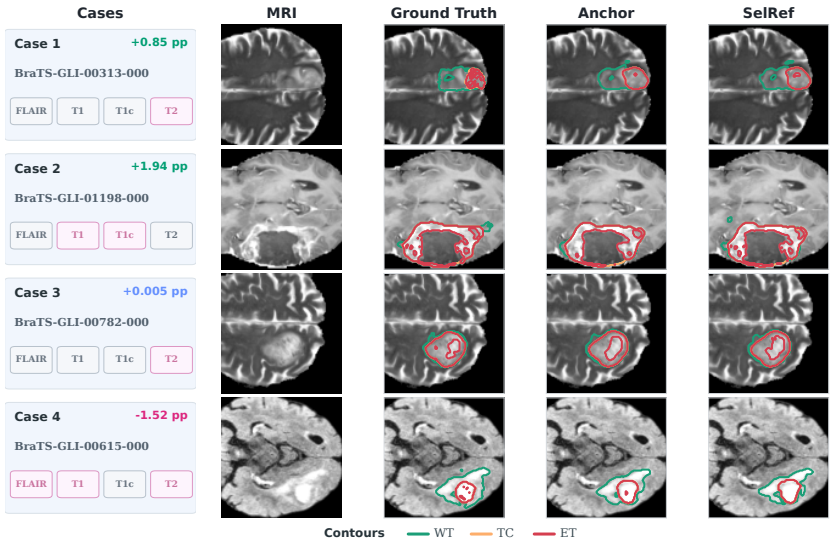

Figure 2: **Qualitative analysis of anchor-relative refinement on BraTS2023.** Green, orange, and red contours indicate WT, TC, and ET regions. Δ Avg. Dice denotes the signed change in Avg. Dice from the prediction anchor to SelRef in percentage points (pp).

Table 4: Computational cost and runtime on BraTS2023.

Method	FLOPs (G) ↓	Params. (M) ↓	Training (s) ↓	Inference (s) ↓
U-HVED	121.601	**1.251**	1.02	0.10
mmFormer	245.529	36.392	1.09	0.19
SFusion	849.973	18.649	1.38	0.59
ShaSpec	1441.978	187.708	0.66	0.14
IMFuse	214.592	47.344	1.46	0.29
SimMLM	269.726	7.816	0.45	0.07
AdaMM	325.881	7.102	0.31	**0.06**
SelRef (Ours)	**6.165**	9.197	**0.27**	0.08

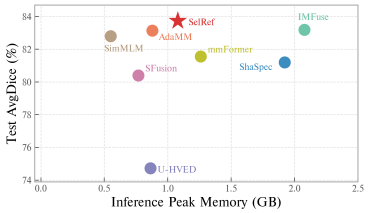

Figure 3: Average Dice versus peak inference memory on BraTS2023.

improves contour alignment with the ground truth, yielding Avg. Dice gains of 0.85 and 1.94 percentage points, respectively. The improvement is particularly visible along the lesion boundary in Case 2. Although Cases 1 and 3 share the same observed modality (T2), Case 3 changes by only 0.005 percentage points, suggesting that refinement depends on the current prediction state beyond the observed-modality configuration. Case 4 presents a failure case, where refinement decreases Avg. Dice by 1.52 percentage points. These examples demonstrate that anchor-relative refinement can improve local predictions while its effectiveness varies across prediction states.

4.5 Model Complexity and Computational Efficiency Analysis

We evaluate the computational efficiency of SelRef in terms of model size, FLOPs, runtime, and inference memory. Table 4 reports the measurements on BraTS2023, and Fig. 3 illustrates the trade-off between segmentation performance and computational cost. All results are obtained with a single 128^3 patch and batch size 1 on an NVIDIA GeForce RTX 3090 using FP16 automatic mixed precision. SelRef introduces limited computational overhead, requiring only 9.197 M parameters and 6.165 G FLOPs. It reduces the computational cost substantially compared with existing baselines, while achieving a training time of 0.27 s and an inference time of 0.08 s, comparable to the most efficient methods. As shown in Fig. 3, SelRef maintains competitive segmentation performance with moderate inference memory consumption. Moreover, the full-modality teacher is involved only during training, and inference requires the same observed modalities available at test time.

AI USE STATEMENT

In this work, artificial intelligence (AI) tools, specifically large language models (LLMs), were used exclusively for language editing to enhance the clarity, readability, and fluency of the manuscript. The authors take full responsibility for all scientific content, including the study design, experimental results, and conclusions. LLMs were not used for research ideation, data analysis, or interpretation of the findings. All AI-assisted text was carefully reviewed and verified by the authors.

REPRODUCIBILITY STATEMENT

The source code, pre-trained models, and configuration files will be made publicly available upon acceptance. All experiments were conducted using publicly available datasets and standard pre-extracted features.

REFERENCES

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *International Conference on Learning Representations (ICLR) Workshop*, 2017. URL https://openreview.net/forum?id=S12_nquOe.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. doi: 10.48550/arXiv.2107.02314.

Ujjwal Baid, Reuben Dorent, Sylwia Malec, Monika Pytlarz, Ruisheng Su, Navodini Wijethilake, Spyridon Bakas, and Alessandro Crimi (eds.). *Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation*, volume 14669 of *Lecture Notes in Computer Science*. Springer, Cham, 2024. ISBN 978-3-031-76163-8. doi: 10.1007/978-3-031-76163-8. MICCAI Challenges, BraTS 2023 and CrossMoDA 2023.

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*, 4:170117, 2017. doi: 10.1038/sdata.2017.117.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*, 2018. doi: 10.48550/arXiv.1811.02629.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Jun-Ho Choi and Jong-Seok Lee. EmbraceNet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270, 2019. doi: 10.1016/j.inffus.2019.02.010.

Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 74–82. Springer International Publishing, 2019. doi: 10.1007/978-3-030-32245-8_9.

Mohammad Hamghalam, Alejandro F. Frangi, Baiying Lei, and Amber L. Simpson. Modality completion via gaussian process prior variational autoencoders for multi-modal glioma segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12907 of *Lecture Notes in Computer Science*, pp. 442–452. Springer Nature, 2021. doi: 10.1007/978-3-030-87234-2_42.

Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. HeMIS: Hetero-modal image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pp. 469–477. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46723-8_54.

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 *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 772–781. Springer International Publishing, 2020. doi: 10.1007/978-3-030-59710-8_75.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12458–12468, 2024. doi: 10.1109/CVPR52733.2024.01184.

Sijie Li, Chen Chen, and Jungong Han. SimMLM: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24068–24077, 2025. doi: 10.1109/ICCV51701.2025.02231.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1657–1665, 2023a. doi: 10.1609/aaai.v37i2.25253.

Zecheng Liu, Jia Wei, Rui Li, and Jianlong Zhou. SFusion: Self-attention based n-to-one multimodal fusion block. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14221 of *Lecture Notes in Computer Science*, pp. 159–169, Cham, 2023b. Springer. doi: 10.1007/978-3-031-43895-0_15.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.

Xiangxi Meng, Kaicong Sun, Jun Xu, Xuming He, and Dinggang Shen. Multi-modal modality-masked diffusion network for brain MRI synthesis with random modality missing. *IEEE Transactions on Medical Imaging*, 43(7):2587–2598, 2024. doi: 10.1109/TMI.2024.3368664.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016. doi: 10.1109/TPAMI.2015.2461544.

Vittorio Pipoli, Alessia Saporita, Kevin Marchesini, Costantino Grana, Elisa Ficarra, and Federico Bolelli. IM-Fuse: A mamba-based fusion block for brain tumor segmentation with incomplete modalities. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15967 of *Lecture Notes in Computer Science*, pp. 225–235. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04984-1_22.

Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and M. Salman Asif. Robust multimodal learning via cross-modal proxy tokens. *Transactions on Machine Learning Research*, oct 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Wtc6wvcYJ0.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(61):2023–2049, 2015.

Hao Wang, Shengda Luo, Guosheng Hu, and Jianguo Zhang. Gradient-guided modality decoupling for missing-modality robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15483–15491, 2024. doi: 10.1609/aaai.v38i14.29474.

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

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 *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14223 of *Lecture Notes in Computer Science*, pp. 216–226. Springer Nature Switzerland, 2023b. doi: 10.1007/978-3-031-43901-8_21.

Shuai Wang, Zipei Yan, Daoan Zhang, Haining Wei, Zhongsen Li, and Rui Li. Prototype knowledge distillation for medical segmentation with missing modality. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023c. doi: 10.1109/ICASSP49357.2023.10095014.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6. IEEE, 2015. doi: 10.1109/ICMEW.2015.7169757.

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

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 438–446. Association for Computing Machinery, 2024. doi: 10.1145/3664647.3681683.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4400–4407. Association for Computing Machinery, 2021. doi: 10.1145/3474085.3475585.

Zhilin Zeng, Zelin Peng, Xiaokang Yang, and Wei Shen. Missing as masking: Arbitrary cross-modal feature reconstruction for incomplete multimodal brain tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15008 of *Lecture Notes in Computer Science*, pp. 424–433. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72111-3_40.

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 *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 107–117. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-16443-9_11.

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*, 112:104108, 2026. doi: 10.1016/j.media.2026.104108.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. CMAD: Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4626–4636, 2025. doi: 10.1109/ICCV51701.2025.00440.

SelRef: Selective Privileged Refinement for Multimodal Learning with Missing Modalities (Appendix)

A Notations 14

B Additional Methodological Details 15

B.1 Task-specific instantiations 15

B.2 Task-aligned coordinates for categorical predictions 15

B.3 Theoretical analysis of selective directional guidance 15

C Experimental Settings 17

C.1 Datasets and evaluation roles 17

C.2 Missing-modality protocol 17

C.3 Evaluation metrics 17

C.4 Implementation details 18

C.5 Baselines 18

D More Experimental Results 20

D.1 BraTS2023 results by missing-modality count 20

D.2 BraTS2020 independent in-domain evaluation 21

D.3 Additional ablation studies 21

E Analysis of Refinement Behavior 23

E.1 Counterintuitive refinement rate 23

E.2 Primary paired results 23

E.3 Results by missing-modality count 23

E.4 Anchor quality and refinement outcomes 24

E.5 Guidance selection across subsets 24

F Robustness under Noisy Modalities 26

G Additional Analyses 27

G.1 Parameter sensitivity analysis 27

G.2 Limitations 27

THE USE OF LARGE LANGUAGE MODELS (LLMS)

Large language models were used to assist with language editing and the presentation of the manuscript. The research questions, methodological development, experimental design, implementation, results, and conclusions were determined and verified by the authors. The authors reviewed all LLM-assisted text and take full responsibility for the content of this work.

A NOTATIONS

Table 5 summarizes the notation used in the main paper and the appendix. Superscript S indicates dependence on the observed modality subset, while subscript u indexes the task unit at which privileged guidance is evaluated.

Table 5: Summary of the main notation.

Notation	Definition
$\mathcal{M} = \{1, \dots, M\}$	Complete set of modalities.
$S \subseteq \mathcal{M}$	Nonempty subset of modalities observed for the current input.
x^m, y	Input from modality m and the corresponding task target.
E_m	Modality-specific expert for modality m.
$\mathbf{h}_m$, $\mathbf{z}_m$	Intermediate representation and native output score produced by E_m.
ρ_τ	Task-dependent map from native scores to predictions.
$\mathbf{p}_m$	Prediction of modality expert m, where $\mathbf{p}_m = \rho_\tau(\mathbf{z}_m)$.
$\mathbf{q}_m$, $\bar{\mathbf{q}}_S$	Compact descriptor of modality m and the pooled descriptor of the observed subset.
α_m^S	Subset-conditioned aggregation weight assigned to modality $m \in S$.
$\mathbf{p}_A^S$, $\mathbf{z}_A^S$	Subset-conditioned prediction anchor and its additive score representation.
ϕ_τ	Task-dependent map from a prediction to an additive score representation.
R_θ	Refinement network conditioned on the observed representations, anchor prediction, and modality subset.
$\boldsymbol{\delta}^S$	Anchor-relative score correction inferred from the observed modalities.
$\mathbf{p}_R^S$, $\mathbf{z}_R^S$	Refined prediction and its additive score representation.
T, $\mathbf{z}_T$	Frozen full-modality teacher and its output score, used during training.
Γ_τ	Fixed map from additive scores to the task-aligned score space.
$\mathbf{s}_A^S$, $\mathbf{s}_R^S$, $\mathbf{s}_T$	Task-aligned scores of the anchor, refined prediction, and teacher.
$\ell_\tau(\mathbf{s}, y)$	Differentiable convex reference loss in the task-aligned score space.
$\mathcal{U}_\tau$, u	Set of task units and an individual task-unit index.
g_u^S	Privileged Task Gain of the teacher relative to the anchor.
$\mathbf{r}_{T,u}^S$, $\mathbf{d}_{T,u}^S$	Anchor-relative teacher displacement and its normalized direction.
$\mathbf{r}_{R,u}^S$, $\mathbf{d}_{R,u}^S$	Anchor-relative refinement displacement and its normalized direction.
c_u^S	Current-State Compatibility between the teacher direction and local descent direction at $\mathbf{s}_{R,u}^S$.
$w_{\mathrm{gain},u}^S$	Detached weight derived from positive Privileged Task Gain.
$w_{\mathrm{cmp},u}^S$	Detached weight derived from positive Current-State Compatibility.
$w_{\mathrm{norm},u}^S$	Detached attenuation factor for small teacher or refinement displacements.
$\ell_{\mathrm{dir},u}^S$	Directional discrepancy between teacher and refinement displacements.
$\mathcal{L}_{\mathrm{task}}$, $\mathcal{L}_{\mathrm{priv}}$	Downstream task loss and selective privileged directional loss.
λ_{priv}	Weight of the privileged directional loss.
τ_g, τ_T, τ_R	Saturation or attenuation scales used by the guidance weights.

Task	Γ_τ	Task unit u	ℓ_τ
Food-101	Centered categorical logits	Example	Multiclass cross-entropy
MM-IMDb	Bernoulli logits (identity)	Example–label pair	Binary cross-entropy
BraTS	WT/TC/ET region logits	Voxel	Mean region-wise binary cross-entropy

Table 6: Task-dependent score spaces and reference losses. The BraTS mapping acts on four-class voxel logits and produces WT, TC, and ET region logits.

B Additional Methodological Details

B.1 Task-specific instantiations

The prediction map ρ_τ, its score representation ϕ_τ, and the task-aligned map Γ_τ are specified according to the output structure. Table 6 identifies the corresponding reference loss and task unit used to assess teacher guidance. The reference loss governs guidance selection; the downstream task loss supervises the refined prediction through Eq. 11.

For Food-101, ρ_τ is softmax and ϕ_τ maps a strictly positive probability vector to centered log-probabilities. For MM-IMDb, ρ_τ is the elementwise sigmoid and ϕ_τ is the elementwise logit. BraTS uses the corresponding voxelwise four-class softmax and centered log-probabilities to form its additive anchor score representation. Probability clipping makes these inverse score maps numerically stable; at clipped values they are approximations.

For BraTS, let $\mathbf{z} = (z_0, z_1, z_2, z_3)$ be the four mutually exclusive voxel logits. The fixed task-aligned transformation is

$$\begin{aligned} s_{\mathrm{WT}} &= \mathrm{LSE}(z_1, z_2, z_3) - z_0, \\ s_{\mathrm{TC}} &= \mathrm{LSE}(z_1, z_3) - \mathrm{LSE}(z_0, z_2), \\ s_{\mathrm{ET}} &= z_3 - \mathrm{LSE}(z_0, z_1, z_2), \end{aligned} \tag{12}$$

where LSE denotes log-sum-exp over its arguments. These scores are the logits of the probabilities of the three tumor regions implied by the four-class softmax. With voxel targets $y_{\mathrm{WT}} = \mathbb{I}[y > 0]$, $y_{\mathrm{TC}} = \mathbb{I}[y \in \{1, 3\}]$, and $y_{\mathrm{ET}} = \mathbb{I}[y = 3]$, the convex reference loss at each voxel is the mean of three binary cross-entropies in these region scores. The mapping is invariant to a common shift of the four input logits. It is fixed and differentiable, although it is nonlinear in the native logits.

B.2 Task-aligned coordinates for categorical predictions

Softmax predictions are invariant to a common logit shift: $\mathrm{Softmax}(\mathbf{z} + a\mathbf{1}) = \mathrm{Softmax}(\mathbf{z})$. Accordingly, raw-logit directions contain a prediction-invariant component parallel to $\mathbf{1}$. The centering operator $\mathbf{P} = \mathbf{I} - \frac{1}{C}\mathbf{1}\mathbf{1}^\top$ is symmetric, idempotent, and satisfies $\mathbf{P}\mathbf{1} = \mathbf{0}$. Thus $\Gamma_\tau(\mathbf{z}) = \mathbf{P}\mathbf{z}$ removes this component and provides a unique representative for directional comparison. For a dense categorical output, the same operator can be applied at each spatial location. The BraTS evaluation instead uses the fixed region transformation in Eq. 12.

B.3 Theoretical analysis of selective directional guidance

We first restate the geometric implication of Privileged Task Gain. By the first-order inequality for differentiable convex functions,

$$\ell_\tau(\mathbf{s}_{T,u}, y_u) \geq \ell_\tau(\mathbf{s}^S_{A,u}, y_u) + \nabla \ell_\tau(\mathbf{s}^S_{A,u}, y_u)^\top (\mathbf{s}_{T,u} - \mathbf{s}^S_{A,u}). \tag{13}$$

When $g^S_u > 0$, the teacher has a lower reference loss than the anchor. The inner product on the right-hand side is therefore negative, so the anchor-relative teacher displacement is a descent direction at $\mathbf{s}^S_{A,u}$.

As the refined prediction moves away from the anchor, Current-State Compatibility evaluates the same teacher direction at $\mathbf{s}^S_{R,u}$. A positive c^S_u gives $\nabla \ell_\tau(\mathbf{s}^S_{R,u}, y_u)^\top \mathbf{d}^S_{T,u} < 0$, which is the descent condition in Proposition 2. The following result further connects the selected teacher direction to the direction produced by the refinement network.

Proposition 3 (Descent Transfer through Directional Alignment). Assume that $\mathbf{d}^S_{T,u}$ and $\mathbf{d}^S_{R,u}$ are unit vectors and $\nabla \ell_\tau(\mathbf{s}^S_{R,u}, y_u) \neq \mathbf{0}$. Let

$$\widehat{\mathbf{v}}^S_{R,u} = -\frac{\nabla \ell_\tau(\mathbf{s}^S_{R,u}, y_u)}{\|\nabla \ell_\tau(\mathbf{s}^S_{R,u}, y_u)\|_2}, \qquad \bar{c}^S_u = \left\langle \widehat{\mathbf{v}}^S_{R,u}, \mathbf{d}^S_{T,u} \right\rangle. \tag{14}$$

For nonzero gradients and teacher displacements, $\bar{c}^S_u$ and the stabilized compatibility score c^S_u have the same sign. If $\bar{c}^S_u > 0$ and

$$\ell^S_{\mathrm{dir},u} = 1 - \left\langle \mathbf{d}^S_{R,u}, \mathbf{d}^S_{T,u} \right\rangle < \frac{(\bar{c}^S_u)^2}{2}, \tag{15}$$

then $\mathbf{d}^S_{R,u}$ is a descent direction of the reference loss at the current refined prediction:

$$\nabla \ell_\tau(\mathbf{s}^S_{R,u}, y_u)^\top \mathbf{d}^S_{R,u} < 0. \tag{16}$$

Proof. Since both directions have unit norm,

$$\|\mathbf{d}^S_{R,u} - \mathbf{d}^S_{T,u}\|_2^2 = 2\ell^S_{\mathrm{dir},u}. \tag{17}$$

Applying the Cauchy–Schwarz inequality yields

$$\left\langle \widehat{\mathbf{v}}^S_{R,u}, \mathbf{d}^S_{R,u} \right\rangle = \bar{c}^S_u + \left\langle \widehat{\mathbf{v}}^S_{R,u}, \mathbf{d}^S_{R,u} - \mathbf{d}^S_{T,u} \right\rangle \tag{18}$$

$$\geq \bar{c}^S_u - \|\mathbf{d}^S_{R,u} - \mathbf{d}^S_{T,u}\|_2 \tag{19}$$

$$= \bar{c}^S_u - \sqrt{2\ell^S_{\mathrm{dir},u}} > 0. \tag{20}$$

Thus, the refinement direction has a positive inner product with the normalized negative gradient, proving the stated descent property. □

When ℓ_τ is additionally L-smooth in a neighborhood of $\mathbf{s}^S_{R,u}$, the descent lemma also provides a finite improving interval. In particular, let $a = -\nabla \ell_\tau(\mathbf{s}^S_{R,u}, y_u)^\top \mathbf{d}^S_{R,u} > 0$. Then every $0 < \eta < 2a/L$ satisfies

$$\ell_\tau(\mathbf{s}^S_{R,u} + \eta \mathbf{d}^S_{R,u}, y_u) < \ell_\tau(\mathbf{s}^S_{R,u}, y_u). \tag{21}$$

This result explains how directional alignment transfers the local descent property of a selected teacher direction to the learned refinement as their angular discrepancy decreases.

The implementation follows the same task-unit granularity as the theoretical construction:

- if $\|\mathbf{r}^S_{T,u}\|_2 = 0$ or $\|\mathbf{r}^S_{R,u}\|_2 = 0$, the corresponding direction is zero and $w^S_{\mathrm{norm},u}$ suppresses directional supervision;
- if $g^S_u \leq 0$ or $c^S_u \leq 0$, the corresponding detached weight is zero;
- the gain, compatibility, and norm weights are detached when optimizing the refinement network;
- selection is performed at the task-unit granularity declared in Table 6; sparse selected units are not renormalized, consistent with Eq. 10.

Together, Privileged Task Gain and Current-State Compatibility identify teacher directions that are locally favorable at the anchor and the current refined prediction. The directional loss encourages the observed-modality refinement to approach these directions, while the downstream task objective determines the correction magnitude.

C Experimental Settings

C.1 Datasets and Evaluation Roles

We evaluate SelRef on brain tumor segmentation and bimodal classification. BraTS2023 serves as the primary segmentation benchmark, while UPMC Food-101 and MM-IMDb examine the same missing-modality setting on single-label and multi-label classification, respectively. The benchmark roles are kept separate: BraTS2023 is the primary segmentation study, whereas the classification benchmarks assess the same missing-modality problem with different prediction structures.

BraTS2023. BraTS2023 (Baid et al., 2021; 2024) provides four MRI modalities: FLAIR, T1, contrast-enhanced T1 (T1ce, denoted T1c in our tables), and T2. We evaluate segmentation of the whole tumor (WT), tumor core (TC), and enhancing tumor (ET). For BraTS2023, we adopt the fixed partition used by IM-Fuse (Pipoli et al., 2025), comprising 875 training, 125 validation, and 251 test cases. MRI preprocessing and augmentation follow the mmFormer protocol (Zhang et al., 2022).

BraTS2020. BraTS2020 (Menze et al., 2015; Bakas et al., 2017; 2018) provides the same four MRI sequences and tumor regions. Its 369 cases are split into 258 training, 37 validation, and 74 test cases. The BraTS2020 models are trained and selected within this dataset, without using BraTS2023 training examples.

UPMC Food-101. UPMC Food-101 (Wang et al., 2015) contains paired food images and text for classification into 101 mutually exclusive categories. Following the preprocessing used by CMPT (Reza et al., 2025), the partition used in our experiments contains 61,177 training, 6,795 validation, and 22,716 test examples.

MM-IMDb. MM-IMDb (Arevalo et al., 2017) pairs movie posters with plot descriptions. Each movie can belong to multiple genres among 23 categories. We use 15,552 training, 2,608 validation, and 7,799 test examples, following the partition used by CMPT (Reza et al., 2025).

C.2 Missing-modality protocol

The four BraTS modalities yield 15 nonempty observed subsets. We evaluate every test subject under each subset, including the complete input, and compute subset-specific metrics before averaging the 15 scores with equal weight. The Food-101 and MM-IMDb evaluations use the three fixed configurations of image-only, text-only, and complete input. These configurations are reported separately; their arithmetic mean is denoted Average in the classification tables. The full-modality teacher supplies privileged signals during refinement training and is absent from the inference path.

C.3 Evaluation Metrics

All results are reported as percentages, with higher values indicating better performance. We compute each metric separately for every observed-modality configuration before aggregating across configurations.

Brain Tumor Segmentation. We report the Dice score for whole tumor (WT), tumor core (TC), and enhancing tumor (ET). For subject i, observed-modality subset S, and tumor region r, let $\widehat{Y}^r_{i,S}$ and Y^r_i denote the predicted and ground-truth voxel sets. The regional Dice score is

$$D^r_{i,S} = \frac{2|\widehat{Y}^r_{i,S} \cap Y^r_i|}{|\widehat{Y}^r_{i,S}| + |Y^r_i|}, \qquad r \in \{\mathrm{WT}, \mathrm{TC}, \mathrm{ET}\}. \tag{22}$$

Empty-region cases are handled consistently by the segmentation evaluator used for all methods. For a fixed subset S, we first average $D^r_{i,S}$ over the N test subjects to obtain $\bar{D}^r_S = N^{-1}\sum_{i=1}^{N} D^r_{i,S}$. The regional Avg. reported in the BraTS comparison tables then assigns equal weight to all 15 nonempty subsets:

$$\mathrm{Avg}^r = \frac{1}{15}\sum_{S\neq\varnothing} \bar{D}^r_S. \tag{23}$$

When a single segmentation score is reported, Avg. Dice is the arithmetic mean of the WT, TC, and ET scores under the same subset aggregation. At the subject–subset level, this corresponds to $(D_{i,S}^{\mathrm{WT}} + D_{i,S}^{\mathrm{TC}} + D_{i,S}^{\mathrm{ET}})/3$, which is also a convenient subject-level summary of the three regions.

For Table 7, the NM$=k$ result averages the subsets containing k missing modalities. These groups contain four, six, four, and one subsets for NM$=3$, NM$=2$, NM$=1$, and Full, respectively. Consequently, the all-subset Avg. weights the four group means by their numbers of subsets, rather than assigning equal weight to the four groups.

UPMC Food-101. Food-101 is a single-label classification task. For each observed-modality subset S, accuracy is the proportion of correct predictions among its N test examples:

$$\mathrm{Acc}_S = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\widehat{y}_{i,S} = y_i]. \tag{24}$$

MM-IMDb. MM-IMDb is a multi-label classification task, for which we report macro-averaged F1 across the 23 genre labels. For each subset S, true positives, false positives, and false negatives are counted over the test set separately for each genre c. The resulting score is

$$\text{F1-Macro}_S = \frac{1}{23}\sum_{c=1}^{23} \frac{2\mathrm{TP}_{c,S}}{2\mathrm{TP}_{c,S} + \mathrm{FP}_{c,S} + \mathrm{FN}_{c,S}}. \tag{25}$$

For both classification benchmarks, the Average column is the arithmetic mean of the image-only, text-only, and complete-input scores. Thus, each modality configuration contributes equally, while F1-Macro is computed from test-set predictions before this three-configuration average is taken.

C.4 Implementation details

Training proceeds in stages. Modality-specific experts first learn from their respective observed inputs. Their predictions are then aggregated by the subset-conditioned anchor. A full-modality teacher is trained for privileged supervision, after which the refinement network is optimized while the experts, anchor components, and teacher remain fixed. Inference uses the experts, anchor, and refinement network for the observed modalities.

The BraTS experiments use their respective dataset partitions and preprocessing specified in Appendix C.1; computational measurements in the main paper use a single 128^3 patch, batch size one, FP16 automatic mixed precision, and an NVIDIA GeForce RTX 3090. The classification experiments evaluate the same three input configurations specified in Appendix C.2.

C.5 Baselines

We compare SelRef with methods that address incomplete multimodal inputs through modality completion, representation learning, fusion, expert adaptation, and knowledge distillation. The brain tumor segmentation comparison includes seven methods. Moreover, four of them are also included in the bimodal classification comparison.

U-HVED. U-HVED (Dorent et al., 2019) employs a hetero-modal variational encoder–decoder to learn a shared representation from the available MRI modalities. It jointly performs modality completion and tumor segmentation, providing a comparison with reconstruction-based handling of incomplete inputs.

mmFormer. mmFormer (Zhang et al., 2022) combines modality-specific encoders with intra-modal and inter-modal Transformers. Its inter-modal modeling captures relationships among the available MRI sequences, while the decoder produces a segmentation from their fused representation.

SFusion. SFusion (Liu et al., 2023b) uses self-attention to model interactions among the available modality features and modal attention to obtain a shared representation. Its fusion block accommodates changes in the number of observed modalities across input configurations.

ShaSpec. ShaSpec (Wang et al., 2023a) learns shared and modality-specific features and combines them for prediction. This design retains information common to the observed modalities alongside features that distinguish each modality.

IM-Fuse. IM-Fuse (Pipoli et al., 2025) introduces a Mamba-based fusion block for brain tumor segmentation with incomplete MRI inputs. It provides a comparison with sequence-model-based feature fusion under varying modality availability.

SimMLM. SimMLM (Li et al., 2025) uses a dynamic mixture of modality experts. A learned gating mechanism adjusts the experts' contributions according to the modalities present in the input, making it particularly relevant to SelRef's subset-conditioned prediction anchor.

AdaMM. AdaMM (Zhu et al., 2026) combines adaptive feature modeling with teacher–student knowledge distillation for incomplete MRI segmentation. Its distillation module transfers structural and textural knowledge to the student, providing a teacher-guided comparison for SelRef's selective use of full-modality supervision.

D More Experimental Results

D.1 BraTS2023 Results by Missing-Modality Count

To examine how segmentation performance varies with modality availability, Table 7 aggregates the BraTS2023 results by the number of missing modalities.

When three modalities are missing, SelRef achieves the highest Dice score in all three regions. It exceeds the strongest competing method for ET (SimMLM) by 1.07 percentage points, for TC (AdaMM) by 0.88 points, and for WT (IM-Fuse) by 1.40 points. The WT margin over the best competing result decreases to 0.44, 0.43, and 0.28 points at NM$=2$, NM$=1$, and full modality availability, respectively. Thus, the WT advantage is most pronounced when only one modality is available, while SelRef remains the best method in every WT group.

The results also differ across regions. SelRef leads in every ET group, although the size of its margin does not change monotonically with modality availability. For TC, it leads at NM$=3$ and with full modalities, but ranks second at NM$=2$ (85.65% versus 85.67% for AdaMM) and NM$=1$ (88.81% versus 88.95% for IM-Fuse). Despite these small differences, SelRef retains the highest TC average across all 15 configurations. These grouped results support the benefit of adapting the prediction to the available modalities, particularly under severe missingness, while showing that the gain depends on the segmentation region.

Region	Method	NM$=3$	NM$=2$	NM$=1$	Full	Avg.
ET	U-HVED (Dorent et al., 2019)	50.29	63.64	74.10	81.95	64.09
	mmFormer (Zhang et al., 2022)	61.79	71.52	78.02	82.90	71.42
	SFusion† (Liu et al., 2023b)	59.40	70.80	77.95	84.00	70.55
	ShaSpec (Wang et al., 2023a)	60.72	70.42	77.11	82.77	70.44
	IM-Fuse (Pipoli et al., 2025)	64.61	<u>74.23</u>	80.24	85.22	74.00
	SimMLM (Li et al., 2025)	<u>65.33</u>	73.25	79.93	85.30	73.72
	AdaMM (Zhu et al., 2026)	65.22	74.01	<u>80.36</u>	<u>85.56</u>	<u>74.13</u>
	SelRef (Ours)	**66.40**	**74.68**	**81.22**	**86.06**	**74.97**
TC	U-HVED (Dorent et al., 2019)	65.55	76.12	83.02	87.43	75.90
	mmFormer (Zhang et al., 2022)	79.01	84.41	87.75	90.34	84.26
	SFusion† (Liu et al., 2023b)	76.35	83.02	86.60	89.50	82.63
	ShaSpec (Wang et al., 2023a)	78.19	84.13	87.64	90.62	83.92
	IM-Fuse (Pipoli et al., 2025)	80.02	85.62	**88.95**	<u>91.54</u>	85.41
	SimMLM (Li et al., 2025)	80.08	84.86	87.48	89.25	84.57
	AdaMM (Zhu et al., 2026)	<u>80.43</u>	**85.67**	88.78	91.50	<u>85.49</u>
	SelRef (Ours)	**81.31**	<u>85.65</u>	<u>88.81</u>	**91.97**	**85.76**
WT	U-HVED (Dorent et al., 2019)	76.98	85.15	88.47	89.81	84.17
	mmFormer (Zhang et al., 2022)	84.62	89.46	91.67	92.61	88.97
	SFusion† (Liu et al., 2023b)	83.05	88.58	91.05	92.20	88.01
	ShaSpec (Wang et al., 2023a)	84.74	89.68	92.00	93.15	89.21
	IM-Fuse (Pipoli et al., 2025)	<u>86.69</u>	<u>90.51</u>	92.29	93.14	<u>90.14</u>
	SimMLM (Li et al., 2025)	86.47	90.46	<u>92.32</u>	<u>93.25</u>	90.08
	AdaMM (Zhu et al., 2026)	86.09	90.13	92.16	93.03	89.79
	SelRef (Ours)	**88.09**	**90.95**	**92.75**	**93.53**	**90.84**

Table 7: **Segmentation results on BraTS2023 aggregated by missing-modality count.** NM denotes the number of missing modalities; NM$=3$, NM$=2$, and NM$=1$ therefore correspond to one, two, and three available modalities, respectively. Each reported value is the mean Dice score (%) over all modality subsets within the corresponding group. Avg. is computed over all 15 nonempty modality subsets. Bold and underlined values indicate the best and second-best results, respectively. The † mark identifies the SFusion results sourced from Pipoli et al.

D.2 BraTS2020: Independent In-Domain Evaluation

We further train and evaluate SelRef on BraTS2020 to examine whether the findings on BraTS2023 extend to a separate benchmark release under the same 15 observed-modality configurations. Table 8 reports the Dice scores for each configuration and tumor region.

On BraTS2020, SelRef obtains the highest average WT Dice of 87.66%, slightly exceeding SimMLM at 87.55%. Its average ET Dice of 64.35% ranks second to SimMLM (64.71%), whereas its TC Dice of 78.75% falls below IM-Fuse (79.36%). The results therefore show a different regional pattern from BraTS2023, where SelRef leads on all three regional averages.

Performance also varies with the observed modality subset. With T1c alone, SelRef achieves the best TC and WT scores, reaching 84.70% and 81.13%, respectively. The corresponding margins over the strongest baselines are 0.95 and 1.83 percentage points. With FLAIR alone, it leads on ET at 50.77%, while its TC score remains below the strongest baseline. These differences suggest that the benefit of refinement depends on both the information retained in the available MRI sequence and the tumor region being segmented. Overall, the independent BraTS2020 evaluation supports the WT improvement observed on BraTS2023, while showing that the ET and TC gains do not carry over uniformly across benchmark releases.

	Model	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	F T1 T1c T2	Avg.
Enhancing Tumor (ET)	U-HVED Dorent et al. (2019)	27.74	18.15	67.15	23.63	33.59	69.98	36.67	67.25	27.31	69.25	70.52	37.03	70.24	69.20	70.85	50.57
	mmFormer Zhang et al. (2022)	44.23	43.29	74.73	46.40	49.28	75.29	50.53	75.23	48.65	75.71	75.57	51.57	75.30	76.00	75.37	62.48
	SFusion Liu et al. (2023b)	43.96	36.34	74.52	43.58	47.58	75.54	48.66	75.38	46.36	75.46	75.39	50.18	75.38	75.76	74.94	61.27
	ShaSpec Wang et al. (2023a)	46.89	42.88	72.84	45.93	49.33	74.08	51.56	73.00	48.01	73.99	74.58	52.04	74.18	74.19	74.45	61.86
	IM-Fuse Pipoli et al. (2025)	45.10	39.40	74.77	**48.28**	45.80	74.97	51.51	76.35	**49.83**	76.15	75.06	51.49	75.48	77.48	75.53	62.48
	SimMLM Li et al. (2025)	49.62	**45.36**	**79.29**	47.88	48.35	**79.49**	50.40	**78.90**	47.65	**79.22**	**78.70**	49.43	**79.28**	**78.58**	**78.49**	**64.71**
	AdaMM Zhu et al. (2026)	47.41	43.12	75.88	47.88	51.42	78.75	53.42	76.22	48.85	75.24	77.65	**54.10**	78.45	75.33	76.02	63.98
	SelRef (Ours)	**50.77**	41.25	78.12	46.97	**52.06**	76.44	**53.70**	77.81	48.34	78.29	77.18	53.75	76.78	77.20	76.54	64.35
Tumor Core (TC)	U-HVED Dorent et al. (2019)	57.31	43.90	73.45	37.34	61.52	76.86	58.95	74.56	50.54	75.38	78.73	61.47	77.74	76.62	78.86	65.55
	mmFormer Zhang et al. (2022)	73.81	**68.42**	82.17	71.59	75.90	84.18	75.72	82.82	73.15	83.70	84.21	76.04	84.15	83.90	84.29	78.94
	SFusion Liu et al. (2023b)	70.77	62.74	81.63	67.69	74.15	83.60	73.87	83.45	72.27	83.29	84.19	75.58	83.52	84.22	84.02	77.67
	ShaSpec Wang et al. (2023a)	73.19	67.18	83.75	69.56	74.73	**85.10**	75.79	83.32	71.77	85.00	**85.19**	75.66	**85.51**	84.58	**84.93**	79.02
	IM-Fuse Pipoli et al. (2025)	73.86	68.02	83.69	71.57	75.03	85.09	75.74	84.17	73.54	**85.13**	84.90	75.38	84.97	**84.71**	84.62	**79.36**
	SimMLM Li et al. (2025)	**77.05**	66.12	68.77	**74.91**	**77.97**	81.33	**78.96**	71.21	**75.30**	78.25	81.02	**79.01**	81.93	77.61	81.48	76.73
	AdaMM Zhu et al. (2026)	70.19	67.10	82.38	73.17	74.24	84.12	75.55	82.57	70.68	83.75	84.14	75.26	83.77	83.30	83.94	78.28
	SelRef (Ours)	72.57	68.06	**84.70**	70.32	74.71	84.82	74.25	**84.29**	71.76	84.55	84.12	74.81	84.53	83.83	83.90	78.75
Whole Tumor (WT)	U-HVED Dorent et al. (2019)	82.97	52.52	63.14	73.34	86.07	85.88	86.40	65.43	79.04	78.77	87.36	87.71	88.14	81.14	88.77	79.11
	mmFormer Zhang et al. (2022)	87.62	77.60	76.61	85.40	89.50	89.92	89.54	80.03	86.77	86.90	90.14	89.66	90.13	87.65	90.20	86.51
	SFusion Liu et al. (2023b)	87.73	72.79	75.58	84.28	89.44	89.22	89.61	79.28	85.78	86.79	90.09	89.97	90.15	87.08	90.44	85.88
	ShaSpec Wang et al. (2023a)	88.92	75.96	78.20	84.46	89.65	**90.27**	**90.12**	80.00	86.05	87.02	90.36	90.26	**90.71**	87.55	**90.81**	86.69
	IM-Fuse Pipoli et al. (2025)	88.79	78.04	78.58	85.88	89.88	89.86	89.87	81.10	87.22	87.74	90.21	90.03	90.32	88.08	90.32	87.06
	SimMLM Li et al. (2025)	88.90	**80.21**	79.30	86.29	**90.05**	90.08	**90.12**	81.89	**87.63**	**88.07**	**90.45**	**90.50**	90.69	**88.46**	90.62	87.55
	AdaMM Zhu et al. (2026)	83.65	75.09	76.61	**88.56**	89.44	89.74	89.81	78.65	85.44	86.54	89.71	89.97	90.22	86.54	90.21	86.01
	SelRef (Ours)	**89.14**	79.87	**81.13**	87.84	89.77	90.06	89.84	**82.35**	87.42	87.80	90.13	90.29	90.45	88.11	90.75	**87.66**

Table 8: **Dice score (%) comparison on BraTS2020** under both complete and incomplete modality settings. Filled modality markers indicate observed modalities, whereas hollow markers indicate missing modalities. Results marked by † are taken from Pipoli et al.(Pipoli et al., 2025). Bold and underlined values indicate the best and second-best results, respectively.

D.3 Additional Ablation Studies

We examine whether the progression from the subset-conditioned anchor to task-trained refinement and privileged guidance also appears in bimodal classification. Tables 9 and 10 report the three input configurations for each variant.

On Food-101, task-only refinement and prediction KD obtain average accuracies of 85.87% and 85.86%, close to the anchor's 85.85%. Directional guidance reaches 86.10%, and Full SelRef reaches 86.28%. The image-only result rises from 77.44% for the anchor to 77.82% for Full SelRef, so the average improvement also appears when one modality is unavailable.

On MM-IMDb, Full SelRef obtains the highest F1-Macro under each input configuration and improves the average from 56.25% for the anchor to 56.64%. Its image-only score increases from 50.51% to 50.94%. Across both tasks, the complete objective outperforms directional guidance without selection. These variants support the joint use of guidance selection; they do not isolate the contribution of the gain and compatibility criteria individually.

Variant	Text	Image	Text–Image	Average
Anchor	85.86	77.44	94.24	85.85
Task-only	85.89	77.37	94.35	85.87
KD	85.81	77.45	94.32	85.86
Direction-only	86.04	77.50	94.77	86.10
Full SelRef	**86.15**	**77.82**	**94.86**	**86.28**

Table 9: Ablation results on UPMC Food-101 (accuracy, %).

Variant	Text	Image	Text–Image	Average
Anchor	58.29	50.51	59.94	56.25
Task-only	58.36	50.50	60.13	56.33
KD	58.42	50.48	60.25	56.38
Direction-only	58.43	50.51	60.17	56.37
Full SelRef	**58.55**	**50.94**	**60.42**	**56.64**

Table 10: Ablation results on MM-IMDb (F1-Macro, %).

Variant	Mean $\overline{\Delta}\uparrow$	Improvement rate $\uparrow$	CRR $\downarrow$	CRR@1pt $\downarrow$	Harm magnitude $\downarrow$
Task-only Refinement	2.15	88.69	11.26	4.65	2.61
Full SelRef	**2.30**	**91.69**	**8.26**	**2.90**	**1.93**

Table 11: Paired refinement outcomes on the BraTS2023 locked test set over 251 subjects and all 15 nonempty modality subsets. Mean $\overline{\Delta}$ and harm magnitude are measured in Dice percentage points; improvement rate and CRR are percentages. CRR@1pt counts decreases exceeding one Dice percentage point. All comparisons use raw, unrounded subject-level Dice.

E Analysis of Refinement Behavior

E.1 Counterintuitive Refinement Rate

Privileged Task Gain and Current-State Compatibility characterize teacher directions through the local geometry of the reference loss. We further examine segmentation outcomes in a paired diagnostic run of the learned anchor-relative corrections. Its aggregate checkpoint scores differ from those in the main ablation table; all comparisons in this section are therefore confined to the paired diagnostic run. In particular, an improvement in average Dice may coexist with decreases for individual subjects and observed-modality subsets.

Let $D^r(\widehat{y}, y)$ denote the raw, unrounded Dice score for $r \in \{\mathrm{WT}, \mathrm{TC}, \mathrm{ET}\}$. For subject i and observed-modality subset S, we measure the change induced by refinement as

$$\Delta^r_{i,S} = D^r(\widehat{y}^S_{R,i}, y_i) - D^r(\widehat{y}^S_{A,i}, y_i), \qquad \overline{\Delta}_{i,S} = \frac{1}{3} \sum_{r \in \{\mathrm{WT},\mathrm{TC},\mathrm{ET}\}} \Delta^r_{i,S}. \tag{26}$$

Thus, $\overline{\Delta}_{i,S}$ compares the refined prediction with its anchor under the same input configuration. For an evaluation set $\mathcal{E}$ of subject–subset pairs, we define the Counterintuitive Refinement Rate as

$$\mathrm{CRR}_\varepsilon(\mathcal{E}) = \frac{1}{|\mathcal{E}|} \sum_{(i,S)\in\mathcal{E}} \mathbb{I}\big[\overline{\Delta}_{i,S} < -\varepsilon\big], \qquad \varepsilon = 10^{-12}. \tag{27}$$

The tolerance accounts for floating-point variation. We additionally report the improvement rate $\Pr(\overline{\Delta}_{i,S} > \varepsilon)$, the rate of decreases exceeding one Dice percentage point $\Pr(\overline{\Delta}_{i,S} < -0.01)$, and the mean decrease conditional on a harmful correction:

$$H_\varepsilon = -\mathbb{E}\big[\overline{\Delta}_{i,S} \mid \overline{\Delta}_{i,S} < -\varepsilon\big]. \tag{28}$$

Each of the 251 test subjects is evaluated under all 15 nonempty modality subsets. Consequently, the mean of $\overline{\Delta}_{i,S}$ over these 3,765 pairs equals the difference between the aggregate raw AvgDice scores of the refined prediction and its corresponding anchor. Confidence intervals are estimated by resampling subjects while retaining all 15 subsets of each sampled subject.

E.2 Primary CRR Results

As shown in Table 11, both variants improve their shared anchor on average, while SelRef achieves a larger mean gain and a lower frequency of harmful corrections. For the diagnostic SelRef checkpoint, the mean raw AvgDice rises from 81.41% to 83.71%, consistent with the 2.30-point mean paired gain in the table. Relative to task-only refinement, SelRef reduces CRR from 11.26% to 8.26%. The paired reduction is 3.00 percentage points, with a 95% confidence interval of [1.94, 4.12] points from 10,000 subject-clustered bootstrap samples. Decreases exceeding one Dice point also occur less frequently, falling from 4.65% to 2.90%. These results extend the aggregate comparison by describing both the average benefit and the case-level risk of refinement.

E.3 Refinement under Different Missing-Modality Settings

We next examine whether the average benefit of refinement and its case-level risk vary with the number of available modalities. Table 12 groups the paired raw AvgDice results by the number

Setting	Variant	Mean $\overline{\Delta}\uparrow$	CRR $\downarrow$	CRR@1pt $\downarrow$
NM$=3$	Task-only Refinement	0.85	21.41	7.47
	Full SelRef	0.84	17.13	5.38
NM$=2$	Task-only Refinement	2.05	9.83	4.12
	Full SelRef	2.20	6.64	2.39
NM$=1$	Task-only Refinement	3.20	5.48	3.19
	Full SelRef	3.44	3.78	1.79
Full	Task-only Refinement	3.71	2.39	2.39
	Full SelRef	4.18	0.40	0.40

Table 12: Refinement outcomes on BraTS2023 grouped by the number of missing modalities. NM$=$ 3, NM$=$ 2, and NM$=$ 1 contain four, six, and four observed-modality subsets, respectively; Full contains one. Mean $\overline{\Delta}$ is the change in raw AvgDice from each variant's corresponding anchor, in percentage points. CRR and CRR@1pt are percentages of subject–subset pairs.

	Q1: lower anchor Dice		Q4: higher anchor Dice	
Setting	Mean $\overline{\Delta}$	CRR	Mean $\overline{\Delta}$	CRR
NM$=3$	1.08	29.48	1.10	3.59
NM$=2$	2.76	13.26	2.42	0.27
NM$=1$	4.52	9.56	3.06	0.00

Table 13: SelRef refinement outcomes by anchor quality within each incomplete-modality group on BraTS2023. Quartiles are formed separately within each NM group using anchor raw AvgDice; Q1 and Q4 denote the lowest and highest quartiles. Mean $\overline{\Delta}$ is reported in Dice percentage points, and CRR in percent.

of missing modalities. Both variants improve their respective anchors on average in every group. With only one modality available (NM$=$ 3), however, the mean gain is relatively small and harmful corrections occur more often than under less severe missingness. SelRef reduces CRR from 21.41% to 17.13% in this setting, while the two variants have nearly identical mean Dice gains. As more modalities become available, the mean gains increase and CRR declines. SelRef maintains a lower CRR than task-only refinement in each group.

The regional results provide a complementary view of this pattern. SelRef has a CRR of 4.94% for WT, 17.58% for TC, and 11.74% for ET, compared with 5.07%, 21.27%, and 13.86% for task-only refinement. Thus, TC remains the region with the highest frequency of Dice decreases, although its CRR is lower under SelRef. The paired reduction is also evident for ET; for WT, the estimated difference is small and its confidence interval includes zero.

E.4 Anchor Quality and Refinement Outcomes

Missing-modality severity does not fully describe the prediction state from which refinement begins. We therefore rank subject–subset pairs by the raw AvgDice of their anchors separately within each missing-modality group and divide them into quartiles. Table 13 compares the lowest-quality quartile (Q1) with the highest-quality quartile (Q4). Under NM$=$ 3, CRR falls from 29.48% in Q1 to 3.59% in Q4. A similar association appears under NM$=$ 2 and NM$=$ 1. The mean refinement gain remains positive in the displayed quartiles, but does not increase consistently with anchor quality. These results indicate that weaker anchors offer opportunities for correction while also carrying greater risk of a harmful update.

E.5 Guidance Selection across Modality Subsets

The two selection criteria assess the teacher direction at different prediction states. Positive Privileged Task Gain identifies task units for which the teacher has a lower reference loss than the subset-conditioned anchor. Current-State Compatibility then evaluates whether the same anchor-relative

Setting	$P(g > 0)$	$P(g > 0, c \leq 0)$	$P(g > 0, c > 0)$	$\bar{w}_{\mathrm{sel}}$
NM$=3$	39.56%	0.027%	39.53%	0.00935
NM$=2$	38.95%	0.056%	38.89%	0.00823
NM$=1$	43.82%	0.078%	43.74%	0.00795
Full	47.34%	0.096%	47.25%	0.00751

Table 14: Retrospective guidance statistics for the diagnostic SelRef model on BraTS2023, grouped by the number of missing modalities. Proportions are computed over the fixed evaluation support; $\bar{w}_{\mathrm{sel}}$ is the mean of $w_{\mathrm{gain}} w_{\mathrm{cmp}} w_{\mathrm{norm}}$ over that support. These task-unit statistics describe the selection criteria and are distinct from the subject–subset Dice outcomes in Table 12.

direction remains aligned with descent at the refined prediction. To characterize their empirical behavior, we apply the training-time criteria retrospectively to the locked BraTS2023 test cases and summarize their values over the fixed evaluation support. This diagnostic uses the full-modality teacher and targets after the predictions have been fixed.

Table 14 shows that positive gain occurs on 38.95–47.34% of support units across the four modality groups. Among positive-gain units, a nonpositive compatibility score is rare: its prevalence ranges from 0.027% under NM$=3$ to 0.096% with complete inputs. Consequently, the proportion satisfying both criteria is close to the positive-gain proportion in every group. The mean selective weight also reflects the continuous gain, compatibility, and norm factors, rather than the binary selection rate alone.

Across all 15 subsets, 40.97% of support units have positive gain and 40.91% satisfy both criteria; the mean selective weight is 0.00841. Hence, positive gain accounts for most of the observed selection, while compatibility excludes a small fraction of positive-gain units in this checkpoint. The weights determine the contribution of task units to directional supervision and should be interpreted alongside the fixed-unit reduction in Eq. 10. The CRR analysis above measures the resulting prediction behavior at the subject–subset level.

F Robustness under Noisy Modalities

We evaluate the trained prediction anchor and SelRef under Gaussian perturbations to the observed MRI modalities. The missing-modality configurations remain fixed, and both predictions are computed from the same corrupted inputs at each severity level. As shown in Table 15, both models' Dice scores decrease as the corruption intensifies. SelRef retains higher WT, TC, and ET Dice than its anchor at every reported level. These results describe the response of the learned refinement to degraded observed inputs.

Corruption Severity	Method	WT	TC	ET
Clean	Prediction Anchor	86.78	83.47	71.39
	SelRef	**90.84**(+4.06)	**85.76**(+2.29)	**74.97**(+3.58)
Mild	Prediction Anchor	86.60	82.76	70.82
	SelRef	**90.55**(+3.95)	**84.32**(+1.56)	**73.52**(+2.70)
Moderate	Prediction Anchor	85.24	79.82	67.84
	SelRef	**89.45**(+4.21)	**81.33**(+1.51)	**70.51**(+2.67)
Severe	Prediction Anchor	81.19	74.01	62.57
	SelRef	**86.54**(+5.35)	**75.67**(+1.66)	**64.92**(+2.35)

Table 15: Segmentation performance (%) under noisy observed modalities on the BraTS2023 test set. Results are averaged over 251 subjects and all 15 nonempty modality subsets. Gaussian noise is independently applied to the observed modalities. Mild, Moderate, and Severe correspond to $\gamma \in \{0.10, 0.25, 0.50\}$ times the foreground intensity standard deviation, respectively.

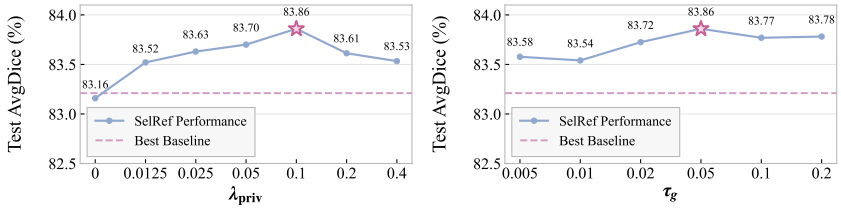

Figure 4: Sensitivity of BraTS2023 average Dice to λ_{priv} and τ_g.

G Additional Analyses

G.1 Parameter Sensitivity Analysis

We investigate two key parameters in SelRef that directly affect privileged guidance on BraTS2023. Figure 4 reports the average Dice when varying each parameter around the default setting.

Balance Between Task Objective and Privileged Guidance. The coefficient λ_{priv} balances the task objective and privileged directional guidance. As shown in Fig. 4, the average Dice increases from 83.16% at $\lambda_{\text{priv}} = 0$ to 83.86% at 0.1, then slightly decreases to 83.61% and 83.53% at 0.2 and 0.4. This trend indicates that appropriate privileged guidance improves refinement, while excessive weighting may weaken the contribution of the task objective.

Robustness to Guidance Selection Scale. The gain-weight scale τ_g controls the contribution of Privileged Task Gain. The performance remains stable across most tested values, ranging from 83.72% to 83.86% for $\tau_g \in [0.02, 0.2]$. The lower results at 0.005 and 0.01 suggest that overly small scales reduce the distinction among teacher guidance with different gains. Overall, the limited variation demonstrates the robustness of SelRef to the choice of τ_g.

G.2 Limitations

SelRef relies on complete-modality cases for training the privileged teacher, which may limit its use when such training cases are scarce. Its theoretical results establish local descent in the task-aligned score space; the finite-step statement additionally assumes local smoothness of the reference loss. Selection uses training targets and does not constitute an inference-time guarantee for individual cases. The observed-modality refinement is therefore assessed by segmentation performance and by the qualitative examples in the main paper, including the reported failure case.