Disentangling for Transfer: Boosting Limited Modalities via Information-Theoretic Regularization and Cross-Modal Reconstruction

Zhiyun Zhang[1,2*], Yan-Jie Zhou[1,3,5*], Yujian Hu[4,6], Xiyao Ma[1,7], Zhouhang Yuan[1,4,5], Zirui Wang[1,3], Hongkun Zhang[6], Minfeng Xu[1,3†]

[1]DAMO Academy, Alibaba Group, Hangzhou, China
[2]Carnegie Mellon University, Pittsburgh, USA
[3]Hupan Laboratory, Hangzhou, China
[4]School of Medicine, Zhejiang University, Hangzhou, China
[5]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[6]Department of Vascular Surgery, The First Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, China
[7]Institute of Automation, Chinese Academy of Sciences, Beijing, China
zhiyunz@andrew.cmu.edu, zhouyanjie.zyj, eric.xmf@alibaba-inc.com

Abstract

Missing critical modalities in medical imaging poses significant challenges for AI-driven diagnostic systems, particularly in scenarios where limited modalities must suffice for downstream tasks. Existing approaches often fail to fully leverage privileged features available only at training or address the information gap between privileged and limited modalities, resulting in suboptimal performance. To address this, we propose a unified, dual-stage **D**isentanglement-**A**lig**N**men**T** fram**E**work (**DANTE**), which uses *Information-Theoretic Regularization* and *Cross-Modal Reconstruction* to decompose full-modality information into alignable and privileged-exclusive components. In the first stage, a self-supervised pre-training strategy based on cross-modal reconstruction acts as a proxy task to implicitly incentivize disentangled representations. In the second stage, we present an information-theoretic regularization to explicitly maximize the transfer of privileged knowledge through two novel modules: (1) a *Mutual Alignment Module* that employs multi-level bidirectional alignment between limited-modality features and alignable features, enhancing cross-modal representation consistency; (2) a *Privileged Compaction Module* that restricts the privileged-exclusive information flow, promoting the integration of task-relevant content into alignable representations. Experimental results on three challenging medical datasets demonstrate that DANTE achieves state-of-the-art performance, demonstrating its effectiveness in leveraging privileged guidance under modality scarcity, and exhibits broad applicability across diverse medical imaging scenarios.

Introduction

Medical imaging is a cornerstone of modern healthcare, providing critical insights for diagnostic tasks such as tumor detection, lesion segmentation, and treatment planning. However, in clinical practice, the absence of certain imaging modalities, especially those that carry essential information for specific downstream tasks, presents significant challenges (Mall et al. 2023). For example, in brain Magnetic Resonance Imaging (MRI), the T1-weighted contrast-enhanced (T1ce) modality is widely regarded as crucial for identifying tumor boundaries and vascular structures due to its ability to highlight areas of blood-brain barrier disruption (Rahimpour et al. 2021). Yet, this modality is often missing in real-world scenarios, either because patients cannot tolerate contrast agents or due to safety concerns such as renal insufficiency (Dill 2008). Similarly, in Computed Tomography (CT) imaging, the absence of computed tomographic pulmonary angiography (CTPA) can obscure vital details about vascular structures and tissue characteristics, limiting the model's ability to make accurate predictions (Wang et al. 2023b). These missing modalities not only reduce the richness of multimodal data but also hinder the performance of AI models, particularly when they rely on features that

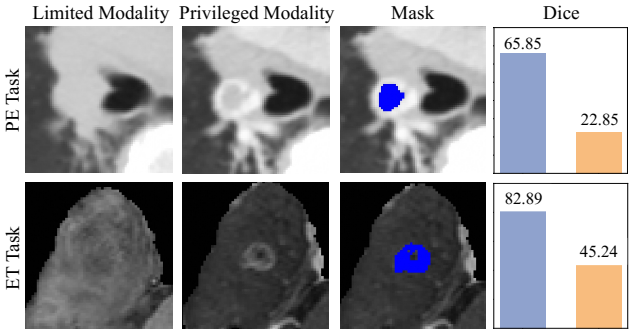

Figure 1: Illustration of the performance gap in limited modality scenarios. Examples shown are: (Top) Pulmonary embolism (PE) task where non-contrast CT is significantly outperformed by the privileged CTPA; (Bottom) Enhancing tumor (ET) task where the limited modality set (e.g., T1, T2, FLAIR) lags behind the privileged T1ce modality. Blue and orange bars denote full and limited modality, respectively.

*These authors contributed equally.
†Hongkun Zhang and Minfeng Xu are corresponding authors.

are uniquely captured by these critical modalities. In many medical scenarios, acquiring complete multimodal data is simply not feasible, necessitating the development of robust limited-modality approaches. However, directly learning from limited data poses additional risks, as models may inadvertently rely on indirect indicators or spurious correlations, leading to biased predictions that fail to generalize well (Izmailov et al. 2022). This raises a fundamental question: how can we fully leverage the dense information from privileged modalities (critical modalities available only in the training data) to significantly enhance the limited modalities' performance at inference? Addressing this challenge is essential for advancing the reliability and applicability of AI-driven medical imaging models in real-world medical settings (Esteva et al. 2019).

To enhance the feature representation capabilities in limited modalities, recent studies have explored various strategies to improve model robustness (Zhang et al. 2022; Chen et al. 2023). However, many existing approaches focus primarily on aligning shared representations across modalities without explicitly supervising the modality-invariant features or constraining the information content of modality-specific features (Havaei et al. 2016; Chen et al. 2019; Wang et al. 2023a). This often results in modality-invariant features that capture only simplistic commonalities, leading to a significant drop in performance when essential modalities are absent. Another aspect of research focuses on cross-modal feature distillation, where knowledge from privileged modalities is transferred to limited ones (Rahimpour et al. 2021; Chen et al. 2021). Due to the substantial information gap between modalities with additional information (e.g., T1ce) and those without, techniques such as projection heads or pooling are often employed before distillation. However, these operations can inadvertently lead to the loss of crucial feature information during the distillation process, hindering the model's ability to enhance limited-modality representations (Rahimpour et al. 2021). Notably, for spatially alignable medical images, the inherent cross-modal correspondences present a significant, yet untapped, opportunity to learn better feature representations. Additionally, reconstruction tasks have been proposed to enhance feature representations by learning more expressive and modality-robust representations. By compelling a shared representation to reconstruct any modality with its corresponding appearance code, the model learns features that are both modality-invariant and semantically faithful. (Chen et al. 2019). However, directly optimizing reconstruction and downstream objectives jointly may lead to conflicting gradients or feature competition, resulting in suboptimal performance on downstream tasks.

Drawing on our analysis, we propose **DANTE** — a unified, dual-stage **D**isentanglement-**A**lig**N**men**T** fram**E**work — to enhance feature representation in limited-modality scenarios. We assume that full-modality features can be decomposed into two components: *alignable* and *privileged-exclusive*. **In the first stage**, we adopt cross-modal self-supervised pre-training that implicitly disentangles features through complementary reconstructions of full-modality data from limited and privileged inputs. **In the second stage**, we present an information-theoretic regularization framework that allows the full-modality pathway to transfer rich, transferable information to the limited-modality pathway without degrading its own performance. The key contributions of our work are summarized as follows:

- The proposed DANTE maximizes the alignment of information between limited-modality and privileged modality representations, enabling the model to fully leverage indirect indicators from limited modalities. Our approach achieves leading performance on two public MRI benchmarks and one in-house CT dataset, demonstrating its effectiveness in handling missing-modality scenarios.
- To address the issue of information loss during transformation functions in cross-modal distillation, we introduce information-theoretic regularization achieved via two novel modules: (1) *Mutual Alignment Module*, which employs multi-level bidirectional alignment to better suit dense prediction tasks and provide acceptable knowledge transfer. (2) *Privileged Compaction Module*, which regularizes privileged-exclusive features to encourage information retention in alignable features.
- To further enhance the feature representation, we propose a self-supervised pre-training strategy based on cross-modal reconstruction, which not only improves modality alignment through a shared feature space but also implicitly disentangles alignable and privileged-exclusive features. This pre-training strategy significantly improves feature robustness under modality absence.
- The designed novel cross-modal pre-training strategy is tailored for dense prediction downstream tasks. Extensive experiments highlight the broad applicability of our approach across various medical imaging scenarios.

Related Works

Incomplete multimodal learning. Recent studies have explored strategies to improve robustness under modality absence (Havaei et al. 2016; Wang and Hong 2023; Qiu et al. 2023). Zhang *et al.* (Zhang et al. 2022) proposed mmFormer, which uses modality-specific encoders for local features and an inter-modal Transformer to model global, cross-modal correlations for brain tumor segmentation. Chen *et al.* (Chen et al. 2023) proposed DFTD, a framework that disentangles images into inter-modality relevant and intra-modality specific features via a region-aware approach. Knowledge is then transferred from a synthesized full-modality view via an imputation-based distillation. However, the existing methods focus on achieving alignment without explicitly supervising the modality-invariant features or constraining the information content of modality-specific features. As a result, the aligned features often capture only simplistic commonalities, which may not sufficiently generalize when essential modalities are absent (Chen et al. 2019; Wang et al. 2023a). This limitation underscores the need for more sophisticated mechanisms to ensure that shared representations are both discriminative and robust.

Privileged knowledge distillation. Another relevant line of research is cross-modal distillation, where knowledge

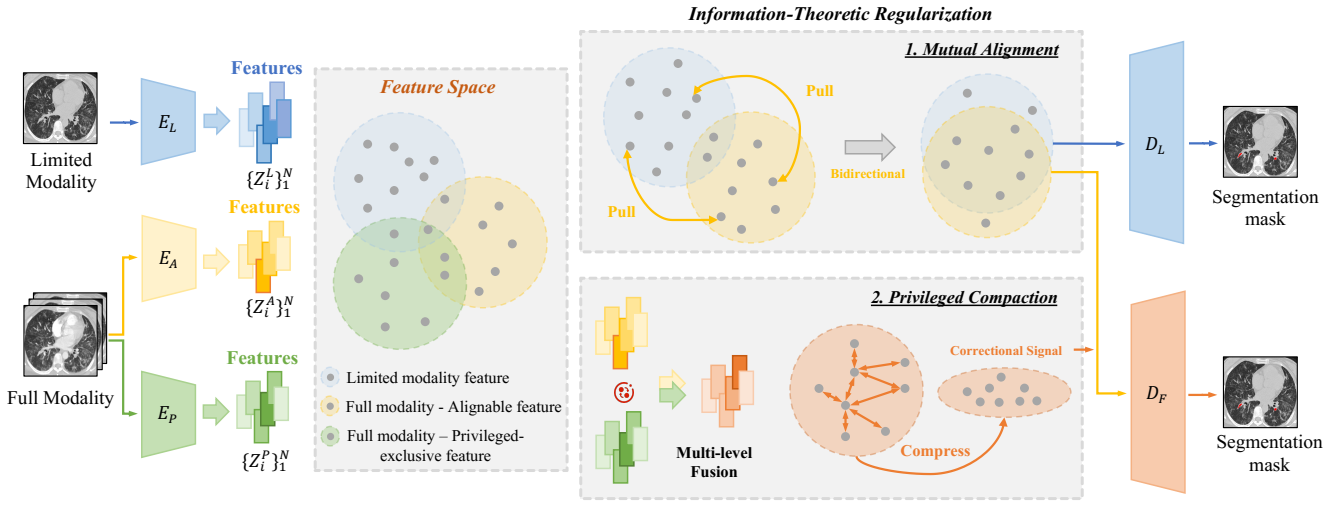

Figure 2: Our proposed Disentanglement-AligNmenT framEwork (DANTE) consists of two pathways: full-modality pathway ($\{E_A, E_P, D_F\}$), available only during training, and limited-modality pathway ($\{E_L, D_L\}$), which is deployed at inference. In the main supervised training stage (Stage 2), we apply information-theoretic regularization to explicitly disentangle full-modality features into alignable feature and privileged-exclusive components for more effective knowledge transfer.

from privileged modalities is transferred to limited ones. These methods typically distill knowledge through logits, feature maps, or relations (Wei, Luo, and Luo 2023). For instance, Rahimpour *et al.* (Rahimpour et al. 2021) distilled logits and features to enable robust segmentation with missing modalities. To mitigate the impact of erroneous teacher predictions, Chen *et al.* (Chen et al. 2021) further introduced a regularized output distillation. Within relation-based methods, Liu *et al.* (Liu et al. 2019) proposed to distill the affinity graph based on pair-wise similarity. More recently, Huo *et al.* (Huo et al. 2024) explored bidirectional distillation to maximize the acceptable knowledge. Despite these advancements, a common limitation persists. These approaches often rely on intermediate operations—such as projection heads or feature summarization operators like pooling and similarity matrices—which can inadvertently discard crucial information. Ultimately, this information loss can constrain the student model's final performance, leaving a critical gap that our work aims to address.

Cross-modal reconstruction. Reconstruction tasks have emerged as a promising avenue to enhance feature representations (Geng et al. 2022; Guo et al. 2024). For dense prediction downstream tasks, Wald *et al.* (Wald et al. 2025) revisited the concept of Mask Autoencoder (MAE) for 3D convolutional networks. In the multi-modal clinical domain, Chen *et al.* (Chen et al. 2019) used shared representations to reconstruct any modality given corresponding appearance codes, thereby improving modality invariance and enriching feature representations. However, simultaneously conducting the reconstruction and downstream tasks faces a fundamental conflict, as optimizing for one can be detrimental to the other. This necessitates a framework that can explicitly reconcile these competing objectives.

Method

Our framework enhances the representation learning of a limited-modality model by leveraging privileged information within a **unified, dual-stage framework** designed to disentangle and transfer knowledge. This is realized across two complementary stages, both built upon a specialized architecture that factorizes representations into *alignable* and *privileged-exclusive* components. We first detail the main supervised training (Stage 2), where an **information-theoretic inspired regularization**, grounded on the Information Bottleneck principle, explicitly steers knowledge transfer. We then describe the foundational pre-training (Stage 1) that makes this process highly effective: a **cross-modal self-supervised learning** that implicitly encourages feature disentanglement through complementary reconstruction objectives, providing a powerful initialization for Stage 2.

Architecture Formulation

Let X^L be the limited modalities, available during both training and inference, and X^P be the privileged modalities. The full-modality input is thus $X^F = \{X^L, X^P\}$, and Y is the target label for the downstream task. Our architecture consists of the following components:

- The **limited-modality model** is denoted as $f_L : X^L \to Y$, where f_L is a composite function of an encoder E_L and a decoder D_L such that $f_L(X^L) = D_L(E_L(X^L))$. This pathway is the one deployed during inference.
- The **full-modality model**, denoted as $f_F : X^F \to Y$. It comprises two specialist encoders, $E_A : X^F \to Z^A$ and $E_P : X^F \to Z^P$, a fusion module ψ, and a decoder D_F. The encoder E_A is trained to extract *alignable information* (Z^A), which ideally represents the maximal information in X^F that is also inferable from X^L, while E_P

extracts *privileged-exclusive information* (Z^P). These representations are fused as $Z^F = \psi(Z^A, Z^P)$ and decoded into the final prediction: $f_F(X^F) = D_F(Z^F)$.

The set of parameters to be optimized is denoted by $\theta = \{\theta_{E_L}, \theta_{E_A}, \theta_{E_P}, \theta_\psi, \theta_{D_L}, \theta_{D_F}\}$.

Information-Theoretic Regularization

Motivated by the Information Bottleneck principle, we formulate an information-theoretic inspired objective comprising two auxiliary regularization modules:

- The **Privilege Compaction** module regularizes the privileged encoder E_P by minimizing the mutual information $I(X^P; Z^P)$. Thus, the full modality model is forced to be "lazy" with privileged features and encode as much task-relevant information as possible into the alignable representation Z^A.
- The **Mutual Alignment** module creates a information-theoretic link between this enriched representation Z^A and the limited-modality representation Z^L by maximizing the mutual information $I(Z^A; Z^L)$.

Combining these elements, our overall *theoretical* objective is to find the optimal set of parameters θ^*:

$$\theta^* = \arg\min_\theta \mathbb{E}_{X,Y}\Big[\underbrace{\mathcal{L}_{\text{Seg}}(f_F(X^F), Y) + \mathcal{L}_{\text{Seg}}(f_L(X^L), Y)}_{\text{Task-Specific Loss}} \underbrace{+I(X^P; Z^P)}_{\text{Privilege Compaction}} \underbrace{-I(Z^A; Z^L)}_{\text{Mutual Alignment}}\Big], \quad (1)$$

where the expectation $\mathbb{E}$ is taken over the training data distribution.

Privilege Compaction. Directly computing and minimizing mutual information $I(X^P; Z^P)$ between high-dimensional variables for a deterministic model is intractable in practice (Saxe et al. 2019). Therefore, we operationalize this principle through a specialized network architecture embodying a structural inductive bias, complemented by a sparsity-based surrogate loss to encourage minimal privileged dependence.

- **Multi-level Residual Fusion Module.** To cater to dense prediction tasks, our encoders E_L, E_A, E_P use hierarchical structures, producing feature maps at N different resolutions. Let $\{Z_i^A\}_{i=1}^N$ and $\{Z_i^P\}_{i=1}^N$ be the sets of multi-level feature maps from encoders E_A and E_P respectively, where i denotes the encoder level. The fusion module ψ operates at each level i with a residual architecture designed to structurally prioritize alignable information:

$$Z_i^F = \psi_i(Z_i^A, Z_i^P) := \text{proj}_i(Z_i^A \oplus Z_i^P) + Z_i^A, \quad (2)$$

where proj_i is a lightweight projection (e.g., 1×1 conv) and $\oplus$ is concatenation. This design establishes Z_i^A as the primary information pathway, while the first term acts as a *correctional signal* that should only carry indispensable privileged information.

- **Privileged-exclusive Information Bottleneck Loss.** Complementing the architecture, we explicitly regularize this correctional signal via the Privileged-exclusive Information Bottleneck (PIB) inspired loss. As minimizing $I(X^P; Z^P)$ can be interpreted as limiting the privileged pathway's information capacity, we adopt L1 penalty as a tractable proxy:

$$\mathcal{L}_{\text{PIB}} = \sum_{i=1}^{N} \left\| \text{proj}_i(Z_i^A \oplus Z_i^P) \right\|_1. \quad (3)$$

Minimizing $\mathcal{L}_{\text{PIB}}$ compels the model to encode only the most salient features in the correctional map, instantiating the $I(X^P; Z^P)$ term in our main objective (Eq. 1).

Mutual Alignment. To promote a high mutual dependency between Z^A and Z^L, we enforce alignment between multi-level feature maps $\{Z_i^A\}_{i=1}^N$ and $\{Z_i^L\}_{i=1}^N$. Our framework's prior feature decoupling makes this dual-alignment strategy viable, mitigating the typical risk of degrading the full-modality pathway. By aligning only the *alignable* component Z^A while preserving the *privileged-exclusive* information within Z^P, we both safeguard the full-modality model's performance and enable it to provide richer contextual guidance to the limited-modality pathway.

We instantiate this objective via two complementary strategies. The first adopts a regression-based formulation that minimizes the ℓ_p-norm distance between corresponding feature maps:

$$\mathcal{L}_{\text{Align}}^{\text{Reg}} = \sum_{i=1}^{N} \left\| Z_i^A - Z_i^L \right\|_p, \quad (4)$$

where $p \in \{1, 2\}$. The second strategy leverages contrastive learning to capture higher-order structural consistency. For each anchor feature vector $z^a \in Z_i^A$, its corresponding feature $z_+^l \in Z_i^L$ is treated as the positive sample, while other features in the batch serve as negatives $\{z_-^l\}$. The total contrastive alignment loss is then formulated as the average pixel-level InfoNCE loss (Oord, Li, and Vinyals 2018) over all anchor vectors $\mathcal{A}$ across all levels:

$$\mathcal{L}_{\text{Align}}^{\text{Contra}} = -\frac{1}{|\mathcal{A}|} \sum_{z^a \in \mathcal{A}} \log \frac{\exp(\text{sim}(z^a, z_+^l)/\tau)}{\sum_{z_j^l \in \{z_+^l\} \cup \{z_-^l\}} \exp(\text{sim}(z^a, z_j^l)/\tau)}, \quad (5)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature hyperparameter. Either $\mathcal{L}_{\text{Align}}^{\text{Reg}}$ or $\mathcal{L}_{\text{Align}}^{\text{Contra}}$ serves as a practical surrogate for the mutual information maximization term $-I(Z^A; Z^L)$ in the overall objective (Eq. 1).

Overall loss for Stage 2. By replacing the intractable mutual information terms in Eq. 1 with practical surrogates, the final objective for the supervised learning stage becomes:

$$\mathcal{L}_{\text{Stage 2}} = \mathcal{L}_{\text{Task-Specific Loss}} + \lambda_1 \mathcal{L}_{\text{PIB}} + \lambda_2 \mathcal{L}_{\text{Align}}, \quad (6)$$

where $\mathcal{L}_{\text{Task-Specific Loss}}$ is the downstream segmentation loss (e.g., Cross-Entropy or Dice), and λ_1, λ_2 are hyperparameters that balance the contributions of each component. We optimize this loss to find the final model parameters θ^*.

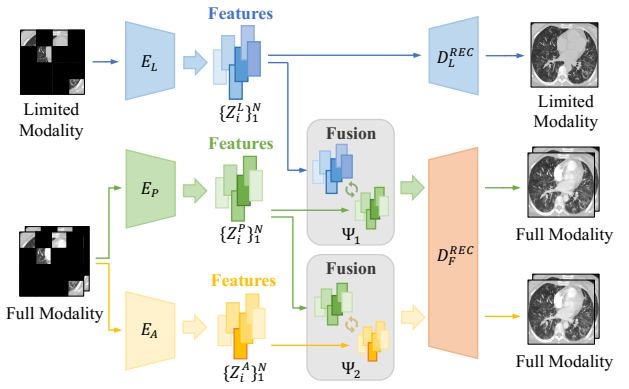

Figure 3: Three cross-modal reconstruction tasks for the self-supervised learning in Stage 1.

Cross-Modal Self-Supervised Learning

To further bolster the feature extraction capabilities of our encoders, we introduce a cross-modal MAE pretraining stage. This stage leverages the main task architecture (E_L, E_A, E_P) for direct parameter transfer and employs two decoders for reconstruction, D_L^{REC} and D_F^{REC}. We adapt a UNet-based convolutional MAE, employing a multi-scale masking strategy suitable for CNNs as proposed in (Wald et al. 2025).

The pretraining is driven by three concurrent reconstruction objectives, formulated as an L2 loss over the masked regions M. First, we establish a baseline by training the limited-modality pathway to reconstruct its own input:

$$\mathcal{L}_{L\to L} = \left\| M \odot (D_L^{\text{REC}}(E_L(X^L)) - X^L) \right\|_2^2. \quad (7)$$

Next, we introduce a pair of reconstruction tasks targeting the full modality X^F via the shared decoder D_F^{REC}. One task combines the alignable and privileged features from X^F:

$$\mathcal{L}_{A+P\to F} = \left\| M \odot (D_F^{\text{REC}}(\Psi_1(E_A(X^F), E_P(X^F))) - X^F) \right\|_2^2. \quad (8)$$

The other, pivotal task reconstructs X^F by fusing limited-modality and privileged features:

$$\mathcal{L}_{L+P\to F} = \left\| M \odot (D_F^{\text{REC}}(\Psi_2(E_L(X^L), E_P(X^F))) - X^F) \right\|_2^2, \quad (9)$$

where Ψ_1, Ψ_2 adopt the same architecture as ψ in Eq. 2, but are independently parameterized. This dual-target design is the key to our implicit regularization. Eq. 7 establishes that Z^L contains sufficient information to reconstruct X^L. Therefore, the complementary task in Eq. 9 encourages Z^P to exclusively encode the information present in X^F but absent in X^L—essentially the *privileged-exclusive* content. Simultaneously, because both $\Psi_1(Z^A, Z^P)$ from Eq. 8 and $\Psi_2(Z^L, Z^P)$ from Eq. 9 must reconstruct the same target X^F via the shared decoder D_F^{REC}, the model is forced to render the alignable features Z^A and Z^L functionally equivalent, driving them into a common latent space.

The total Stage 1 objective is the sum of these losses, $\mathcal{L}_{\text{Stage 1}} = \mathcal{L}_{L\to L} + \mathcal{L}_{A+P\to F} + \mathcal{L}_{L+P\to F}$. The resulting encoder weights are then transferred to initialize the main supervised framework.

Experiments

Experiments Setting

Datasets. We evaluated our method on two public benchmarks and one in-house dataset. All data was split into 70% training, 10% validation, and 20% test sets. We designate the contrast-enhanced scan as the critical *privileged modality*: it is vital for treatment planning yet often absent due to cost and safety risks (Rahimpour et al. 2021; Bai et al. 2024). **(1) Public Benchmarks (BraTS 2018/2020):** The BraTS 2018 and 2020 datasets (Menze et al. 2014) provide 285 and 369 multi-modal MRI cases, respectively, each including paired T1, T2, T1ce, and Flair scans. The segmentation task involves three nested sub-regions: the enhancing tumor (ET), the tumor core (TC), and the whole tumor (WT). T1ce is critical for delineating the Enhancing Tumor (ET), a primary target for surgery and radiotherapy. Therefore, we designate it as the privileged modality. For the cross-modal pre-training stage, we utilize the larger BraTS 2021 dataset, which contains 1,251 cases. **(2) In-House Dataset:** Our in-house Dual-Phase Pulmonary Embolism (DP2E) dataset provides 776 paired CTPA and Non-contrast CT (NCT) scans. **CTPA** is the privileged modality as it is the gold standard for visualizing filling defects, though frequently omitted in emergencies. We supplement pre-training with an additional 2,238 normal lung CT pairs.

Implementation Details. All experiments are conducted on a single NVIDIA H20 GPU. Models are trained for 500 epochs using the SGD optimizer with a batch size of 2, a Nesterov momentum of 0.99, and a weight decay of 3×10^{-5}. We employ a "poly" learning rate schedule with an initial learning rate of 10^{-2}. For the BraTS datasets, we use a patch size of $128 \times 128 \times 128$ with an isotropic spacing of 1.0 mm. For the DP2E dataset, the patch size is $64 \times 160 \times 192$ with a spacing of $2.0 \times 0.7 \times 0.7$ mm. We utilize the standard nnUNetv2 data augmentation suite, which includes spatial, intensity, and mirroring transformations. During inference, a sliding window approach with a 50% overlap and Gaussian importance weighting is employed. For the Stage 2 training, we transfer the pre-trained encoder weights (E_L, E_A, E_P) from Stage 1 and keep them frozen for the first 50 epochs to stabilize training. The privileged compaction and mutual alignment losses are incorporated into the total loss function starting from epoch 100, with their weights set to 0.3 and 0.5, respectively. Unless otherwise specified, we use the L1 norm for the mutual alignment loss ($\mathcal{L}_{\text{Align}}$) due to its empirical stability. A detailed ablation study comparing different alignment strategies ($\mathcal{L}_{\text{Align}}^{\text{Reg}}$ and $\mathcal{L}_{\text{Align}}^{\text{Contra}}$) is provided in the Supplementary Material.

Performance Comparison

Comparative Methods. Firstly, we establish upper and lower bounds by evaluating a plain convolutional UNet in full-modality and critical modality-missing (T1ce or CTPA) settings. Secondly, for state-of-the-art methods, we compare

Model	Enhancing Tumor (ET)		Tumor Core (TC)		Whole Tumor (WT)		DP2E Dice↑
	BraTS 2018	BraTS 2020	BraTS 2018	BraTS 2020	BraTS 2018	BraTS 2020	
Lower Bound	45.24 / 9.23	52.44 / 8.97	68.27 / 5.48	70.51 / 5.31	90.87 / 5.72	91.12 / 6.03	22.85
Upper Bound	82.89 / 2.67	83.64 / 4.17	82.12 / 4.55	84.77 / 5.62	91.20 / 4.95	91.92 / 5.19	65.85
Missing Modality-based							
RobustSeg (TMI'19)	48.84 / 9.50	55.80 / 8.57	71.59 / 5.69	76.24 / 5.93	90.44 / 6.21	91.46 / 5.11	25.67
mmFormer (MICCAI'22)	50.02 / 10.28	56.77 / 9.09	72.97 / 5.72	77.12 / 5.68	90.60 / 5.06	91.53 / 4.83	24.72
ShaSpec (CVPR'23)	48.50 / 10.76	53.06 / 9.29	71.18 / 7.27	74.48 / 6.00	88.76 / 7.29	90.59 / 5.32	24.74
MMANet (CVPR'23)	51.02 / 9.30	57.58 / 8.20	73.18 / 4.85	77.04 / 5.55	91.68 / 4.90	91.73 / 4.67	27.82
Distillation-based							
PMKL (TMI'21)	49.76 / 10.54	58.08 / 8.39	72.96 / 5.48	77.14 / 5.67	91.41 / 5.06	91.73 / 4.79	27.69
CMFD (TMI'22)	51.23 / 10.41	53.59 / 9.30	72.61 / 4.76	74.84 / 5.27	91.55 / 4.18	91.52 / 4.49	28.49
CPMN (MICCAI'24)	49.78 / 11.21	58.11 / 8.18	72.83 / 5.76	77.24 / 5.69	90.96 / 5.32	91.95 / 4.77	29.33
DANTE w/o Stage 1	54.93 / 8.29	61.00 / 8.05	74.83 / **4.19**	78.42 / 5.61	91.87 / **3.67**	92.00 / 4.72	30.10
DANTE	**57.09 / 7.60**	**61.81 / 6.98**	**77.24** / 4.53	**78.89 / 5.22**	**91.99** / 3.98	**92.12 / 4.39**	**31.06**

Table 1: Comparison with state-of-the-art methods on BraTS 2018, BraTS 2020, and DP2E datasets. For BraTS, results are reported as Dice Score (%) (↑) / HD95 distance (↓). DANTE denotes our dual-stage framework, and 'w/o' stands for 'without'.

our model against two categories of approaches: (1) Missing modality-based methods: RobustSeg (Chen et al. 2019), mmFormer (Zhang et al. 2022), ShaSpec (Wang et al. 2023a) and MMANet (Wei, Luo, and Luo 2023). These methods typically focus on extracting modality-invariant features or robustly handling arbitrary modality combinations. (2) Distillation-based methods focus on feature, logit, or relation distillation. This category includes prominent works such as CMFD (Rahimpour et al. 2021), PMKL (Chen et al. 2021), and CPMN (Bai et al. 2024).

To ensure a fair comparison, we mitigate the influence of disparate training strategies and data processing pipelines by integrating the official implementations of all baseline models into the nnUNetv2 framework (Isensee et al. 2021). This approach guaranties that all models are trained and evaluated under identical settings, encompassing data pre-processing, augmentation, training schedules, and inference procedures. For the missing modality-based methods, we employ their designated training strategy, where either the T1ce or CTPA modality is randomly dropped during training.

Results. We compare our proposed DANTE framework against state-of-the-art methods, with quantitative results presented in Table 1. The performance of the *Lower* and *Upper Bounds* first reveals that the absence of critical modalities induces a significant performance gap across both BraTS (particularly for the ET region) and DP2E datasets. Against this backdrop, our DANTE framework, even without pre-training, establishes a strong new baseline. On the BraTS 2018 dataset, it achieves an ET Dice of 54.93%. This not only surpasses the best prior distillation-based method, CMFD (51.23%), by a significant margin of +3.7%, but more importantly, it substantially outperforms the *Lower Bound* by +9.69%. This pattern of superiority is consistent across all tumor sub-regions and datasets. Next, we assess the full dual-stage model, which initializes the encoders using our Stage 1 cross-modal pre-training. This full model further improves performance across all metrics. For instance, DANTE increases the ET Dice on BraTS 2018 to 57.09% and achieves the top score of 31.06% on the DP2E dataset. This indicates that the implicit disentanglement learned during Stage 1 provides a superior feature initialization, enabling the Stage 2 regularization to converge to a more optimal solution. Collectively, these results validate our dual-stage design: Stage 2 provides a powerful supervised regularization framework, while Stage 1 offers a complementary and effective pre-training strategy that further enhances feature quality for the downstream task. Visualization results of segmentation by different methods are provided in the Supplementary Material.

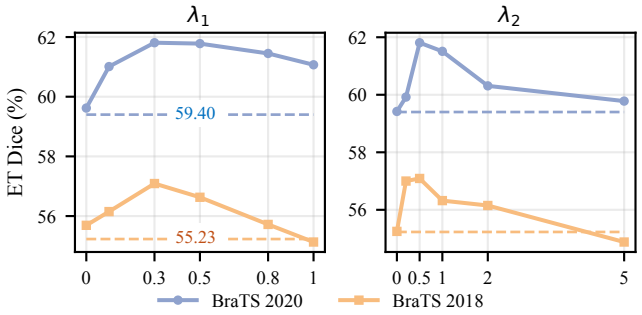

Figure 4: Ablation study on the weights of $\mathcal{L}_{\text{PIB}}$ (λ_1) and $\mathcal{L}_{\text{Align}}$ (λ_2). Dashed lines denote the baseline performance obtained when both losses are disabled ($\lambda_1 = \lambda_2 = 0$). (Left) Effect of varying λ_1 with λ_2 fixed at its optimal value of 0.5. (Right) Effect of varying λ_2 with λ_1 fixed at its optimal value of 0.3.

Analysis

Ablation Study on Privilege Compaction and Mutual Alignment. To evaluate the effectiveness of our proposed $\mathcal{L}_{\text{PIB}}$ and $\mathcal{L}_{\text{Align}}$, we perform an ablation study on their respective weights, λ_1 and λ_2, on both BraTS 2020 and BraTS

Method Configuration	Dataset	Enhancing Tumor (ET)		Tumor Core (TC)		Whole Tumor (WT)	
		Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓
ViT w/o Pretrain	BraTS 2018	40.73	15.31	64.17	15.33	89.59	7.56
	BraTS 2020	43.97	13.01	67.00	12.31	90.38	7.21
ViT w/ Pretrain	BraTS 2018	42.02	13.65	65.68	12.66	90.08	7.19
	BraTS 2020	45.15	12.04	67.93	11.87	90.43	6.00
DANTE w/o Pretrain	BraTS 2018	54.93	8.29	74.83	**4.19**	91.87	**3.67**
	BraTS 2020	61.00	8.05	78.42	5.61	92.00	4.72
DANTE w/o Cross-modal Pretrain	BraTS 2018	55.26	8.09	76.81	4.57	91.92	3.72
	BraTS 2020	60.41	7.68	78.44	5.36	92.08	4.81
DANTE w/ Cross-modal Pretrain	BraTS 2018	**57.09**	**7.60**	**77.24**	4.53	**91.99**	3.98
	BraTS 2020	**61.81**	**6.98**	**78.89**	**5.22**	**92.12**	**4.39**

Table 2: Ablation study of different model configurations and pre-training strategies. 'w/' stands for 'with', 'w/o' for 'without'.

2018 datasets. The optimal weights were determined via a grid search, which yielded $\lambda_1 = 0.3$ and $\lambda_2 = 0.5$. Figure 4 illustrates the impact on the ET Dice score when varying one weight while keeping the other fixed at its optimal value. The dashed lines represent the baseline performance where both weights are set to zero. As shown in Figure 4 (left), with λ_2 fixed at 0.5, the introduction of the $\mathcal{L}_{\text{PIB}}$ boosts performance, peaking at $\lambda_1 = 0.3$. This peak represents a 2.2% (BraTS 2020) and 1.4% (BraTS 2018) improvement over the result at $\lambda_1 = 0$. Similarly, Figure 4 (right) shows that with λ_1 fixed at 0.3, the model performance peaks at $\lambda_2 = 0.5$, achieving a gain of 2.4% and 1.8% compared to when $\lambda_2 = 0$. Notably, performance on both tasks declines as the weights increase past their optimal points. This indicates that an excessive emphasis on these auxiliary objectives can interfere with the primary segmentation task, highlighting the importance of a well-balanced loss function.

Ablation Study on Cross-Modal Self-Supervised Learning. Our ablation study in Table 2 dissects the contributions of our architectural choices and the proposed pre-training strategy. First, we validate our choice of a UNet-based convolutional backbone over a standard Vision Transformer (ViT) for dense prediction tasks. Our DANTE architecture, even trained from scratch, vastly outperforms a pre-trained ViT baseline (Zhou et al. 2023) (e.g., **+15.85%** ET Dice on BraTS 2020), confirming our backbone's suitability. Next, we isolate the impact of our proposed cross-modal pre-training. We observe that applying a naive MAE pre-training ('DANTE w/o Cross-modal Pretrain'), where each modality only reconstructs itself, yields marginal improvements over training from scratch. In contrast, our full cross-modal pre-training yields significant and consistent gains (e.g., ET Dice on BraTS 2018 improves from 54.26% to **57.09%**). This suggests that the performance gains are largely driven by our cross-modal pre-training, rather than by generic pre-training or the architectural choice alone.

Visualization of Feature Space. Figure 5 provides compelling validation of our framework's disentanglement and alignment capabilities. Initially, when pre-trained without our cross-modal reconstruction objective (left), the feature spaces remain completely segregated. In contrast, our Stage 1 pre-training (center) successfully addresses this by implicitly co-locating the Limited-modality and Alignable features into shared clusters, while effectively isolating the Privileged-exclusive information. Finally, our Stage 2 information-theoretic regularization (right) perfects this process: the alignment between the alignable and privileged-exclusive features becomes nearly indistinguishable, and the privileged feature clusters are rendered visibly more compact. This visual progression confirms that our method successfully achieves both feature alignment and privilege compaction, thereby enabling effective knowledge transfer.

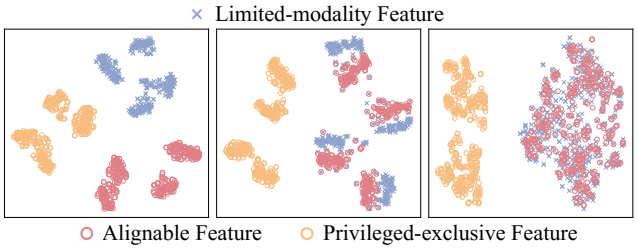

Figure 5: T-SNE visualization of the feature space evolution: (Left) pre-training without cross-modal reconstruction, (Center) with Stage 1 cross-modal pre-training, and (Right) with Stage 2 information-theoretic regularization.

Conclusion

In this work, a novel unified, dual-stage disentanglement-alignment framework (DANTE) is proposed to address the significant challenge of missing critical modalities in medical imaging. In the experiments, our framework achieves state-of-the-art performance on two public MRI benchmarks and one in-house CT dataset, demonstrating its effectiveness in leveraging indirect indicators from limited modalities. More importantly, our proposed approach can seamlessly integrate a novel cross-modality pre-training strategy tailored for dense prediction downstream tasks.

Acknowledgements

This study was supported by the Technical Innovation Key Project of Zhejiang Province (2024C03023) to H.Z.

References

Bai, B.; Zhou, Y.-J.; Hu, Y.; Mok, T. C.; Xiang, Y.; Lu, L.; Zhang, H.; and Xu, M. 2024. Cross-Phase Mutual Learning Framework for Pulmonary Embolism Identification on Non-contrast CT Scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 493–503. Springer.

Chen, C.; Dou, Q.; Jin, Y.; Chen, H.; Qin, J.; and Heng, P.-A. 2019. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 447–456. Springer.

Chen, C.; Dou, Q.; Jin, Y.; Liu, Q.; and Heng, P. A. 2021. Learning with privileged multimodal knowledge for unimodal segmentation. *IEEE Transactions on Medical Imaging*, 41(3): 621–632.

Chen, Y.; Pan, Y.; Xia, Y.; and Yuan, Y. 2023. Disentangle first, then distill: a unified framework for missing modality imputation and Alzheimer's disease diagnosis. *IEEE Transactions on Medical Imaging*, 42(12): 3566–3578.

Dill, T. 2008. Contraindications to magnetic resonance imaging. *Heart*, 94(7): 943–948.

Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; and Dean, J. 2019. A guide to deep learning in healthcare. *Nature Medicine*, 25(1): 24–29.

Geng, X.; Liu, H.; Lee, L.; Schuurmans, D.; Levine, S.; and Abbeel, P. 2022. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*.

Guo, Y.; Sun, S.; Ma, S.; Zheng, K.; Bao, X.; Ma, S.; Zou, W.; and Zheng, Y. 2024. Crossmae: Cross-modality masked autoencoders for region-aware audio-visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26721–26731.

Havaei, M.; Guizard, N.; Chapados, N.; and Bengio, Y. 2016. Hemis: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 469–477. Springer.

Huo, F.; Xu, W.; Guo, J.; Wang, H.; and Guo, S. 2024. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16006–16015.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Izmailov, P.; Kirichenko, P.; Gruver, N.; and Wilson, A. G. 2022. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35: 38516–38532.

Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; and Wang, J. 2019. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2604–2613.

Mall, P. K.; Singh, P. K.; Srivastav, S.; Narayan, V.; Paprzycki, M.; Jaworska, T.; and Ganzha, M. 2023. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics*, 4: 100216.

Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. 2014. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34(10): 1993–2024.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Qiu, Y.; Zhao, Z.; Yao, H.; Chen, D.; and Wang, Z. 2023. Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3228–3239.

Rahimpour, M.; Bertels, J.; Radwan, A.; Vandermeulen, H.; Sunaert, S.; Vandermeulen, D.; Maes, F.; Goffin, K.; and Koole, M. 2021. Cross-modal distillation to improve MRI-based brain tumor segmentation with missing MRI sequences. *IEEE Transactions on Biomedical Engineering*, 69(7): 2153–2164.

Saxe, A. M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B. D.; and Cox, D. D. 2019. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12): 124020.

Wald, T.; Ulrich, C.; Lukyanenko, S.; Goncharov, A.; Paderno, A.; Miller, M.; Maerkisch, L.; Jaeger, P.; and Maier-Hein, K. 2025. Revisiting MAE pre-training for 3D medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5186–5196.

Wang, H.; Chen, Y.; Ma, C.; Avery, J.; Hull, L.; and Carneiro, G. 2023a. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15878–15887.

Wang, M.; Qi, S.; Wu, Y.; Sun, Y.; Chang, R.; Pang, H.; and Qian, W. 2023b. CE-NC-VesselSegNet: Supervised by contrast-enhanced CT images but utilized to segment pulmonary vessels from non-contrast-enhanced CT images. *Biomedical Signal Processing and Control*, 82: 104565.

Wang, Z.; and Hong, Y. 2023. A2fseg: Adaptive multi-modal fusion network for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 673–681. Springer.

Wei, S.; Luo, C.; and Luo, Y. 2023. 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*, 20039–20049.

Zhang, Y.; He, N.; Yang, J.; Li, Y.; Wei, D.; Huang, Y.; Zhang, Y.; He, Z.; and Zheng, Y. 2022. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 107–117. Springer.

Zhou, L.; Liu, H.; Bae, J.; He, J.; Samaras, D.; and Prasanna, P. 2023. Self pre-training with masked autoencoders for medical image classification and segmentation. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)*, 1–6. IEEE.