KiD-Seg: Kinetic-Disentangled Contrastive Learning with Anchor Guidance for Incomplete Multi-Modal Liver Tumor Segmentation

Hee Guan Khor[1,2*(✉)][0000−0002−4935−7815], Dong Zhang[1,3*][0000−0002−2948−1384], Siyan Guo[1][0009−0006−7788−8214], Tianhao Chu[4][0000−0002−4978−7604], Junyi Wang[1][0000−0001−8600−5512], Xiaoyu Bai[1][0009−0001−8070−4707], Yinghong Shi[4][0000−0002−1833−8988], and Le Lu[1][0000−0002−6799−9416]

[1] Ant Group, Hangzhou, Zhejiang, China
xuxichiy19@tsinghua.org.cn
[2] School of Biomedical Engineering, Tsinghua Medicine, Tsinghua University, Beijing, China
[3] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada
[4] Zhongshan Hospital, Fudan University, Shanghai, China
shi.yinghong@zs-hospital.sh.cn

Abstract. Multi-modal MRI is essential for liver tumor delineation but is often incomplete owing to protocol heterogeneity and structured group-wise modality missingness. We address this setting under a clinically grounded acquisition model, where T2-weighted imaging (`T2WI`) serves as a permanent anatomical anchor, while dynamic contrast-enhanced phases (pre-contrast `C-pre`, arterial `C+A`, venous `C+V`, and delayed `C+Delay`), diffusion-weighted imaging (`DWI`), and Dixon sequences (in-phase `InPhase` and out-of-phase `OutPhase`) are available as atomic modality groups. To achieve robust segmentation across all valid permutations, we propose **KiD-Seg** (*Kinetic-Disentangled Contrastive Learning*), a unified framework with three key components: (1) *Asymmetric Anchor-Guided Contrastive Learning*, which aligns auxiliary modalities to the high-fidelity `T2WI` embedding to reduce latent degradation from distortion-prone inputs; (2) *Differential Kinetic Disentanglement*, which encodes phase-to-baseline hemodynamic differences to enhance tumor–background separability; and (3) *Hierarchical Atomic-Group Fusion* with completeness-aware self-distillation, which enforces semantic consistency across sparse and complete modality sets. On the 498-case LLD-MMRI dataset (8 modalities/case), KiD-Seg achieves the best overall median DSC (0.781) and the best full-modality median DSC (0.802), demonstrating robust segmentation performance under both incomplete and complete multi-modal inputs. The source code is publicly available at `https://github.com/AQ-MedAI/KiD_Seg`.

[*] Hee Guan Khor and Dong Zhang contributed equally to this work.

Keywords: Incomplete multi-modal MRI · Liver tumor segmentation · Contrast kinetics · Contrastive learning · Disentanglement.

1 Introduction

Accurate liver tumor delineation supports treatment planning and response assessment, while multi-modal MRI enhances lesion conspicuity and characterization through complementary contrasts [10,15,5]. These include structural anatomy from T2-weighted imaging (`T2WI`), cellularity from diffusion-weighted imaging (`DWI`), fat–water characterization from Dixon imaging (`InPhase`/`OutPhase`), and phase-specific enhancement kinetics from dynamic contrast-enhanced MRI (comprising pre-contrast `C-pre`, arterial `C+A`, portal venous `C+V`, and delayed `C+Delay` phases). Motivated by this clinical setting, we investigate robust liver tumor segmentation on LLD-MMRI [11], comprising eight MRI modalities and seven lesion categories, by exploiting cross-modal complementarity.

Clinical multi-modal MRI often exhibits structured group-wise missingness [7], inducing distribution shifts that impair symmetric fusion by entangling modality-specific artifacts with shared anatomy [2]. While prior works have explored imputation, modality dropout [16,14,12], and disentangled contrastive learning [17,4,3,8] to isolate invariant representations, critical gaps persist for incomplete liver MRI: (i) *symmetric alignment* contaminates high-fidelity anatomical features by pulling them toward distortion-prone sequences (e.g., `DWI`); (ii) temporally coupled dynamic contrast-enhanced (DCE) phases are treated as independent appearances, discarding clinically discriminative kinetic evolution; and (iii) standard independent and identically distributed dropout emulates clinically infeasible scenarios (e.g., isolated DCE phases), diverting capacity and reducing robustness to real-world block-wise protocol failures.

To address these limitations, we propose **KiD-Seg** (*Kinetic-Disentangled Contrastive Learning*), tailored for the group-wise incompleteness in multi-modal liver tumor segmentation. Reflecting clinical protocols, KiD-Seg treats `T2WI` as an immutable reference and groups auxiliary modalities into atomic blocks. First, we propose **Asymmetric Anchor-Guided Contrastive Learning** (AGCL), which establishes T2WI as a robust structural reference. By enforcing a unidirectional alignment of auxiliary modalities toward this anchor, AGCL insulates the shared anatomical representation from the noise of distortion-prone sequences. Second, **Differential Kinetic Disentanglement** (DKD) encodes phase-to-baseline difference maps to explicitly capture local hemodynamic trajectories, utilizing a kinetic contrastive objective to maximize tumor-background separability. Finally, **Hierarchical Atomic-Group Fusion** (HGF) combines structured block-wise dropout with completeness-aware self-distillation, enforcing semantic consistency from minimal (`T2WI`-only) to complete inputs (8 modalities). Extensive validation on the LLD-MMRI dataset [11] strongly substantiates that **KiD-Seg** achieves the best overall robustness and remains competitive across all clinically plausible patterns of modality absence.

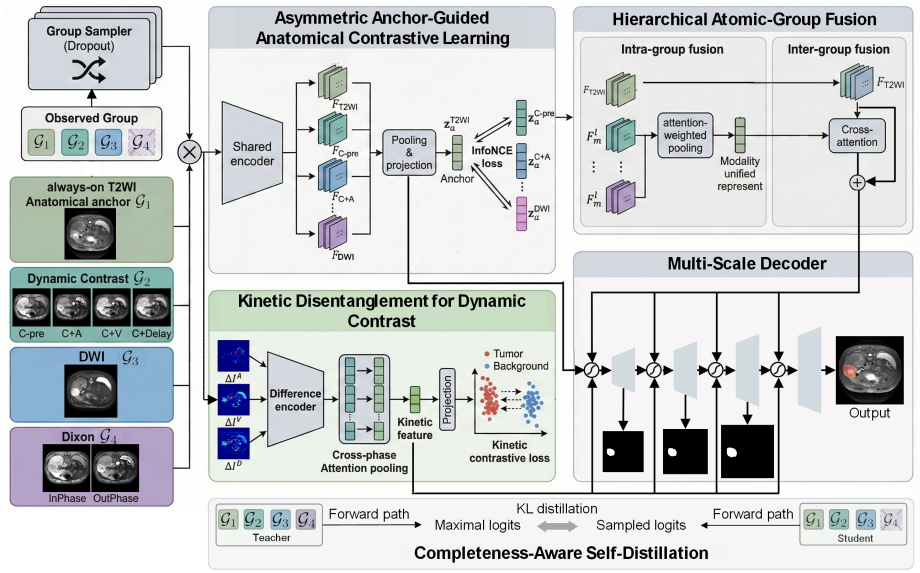

Fig. 1. Architecture of KiD-Seg. `T2WI`-anchored asymmetric contrastive learning, kinetic disentanglement, hierarchical fusion, gated decoding, and completeness-aware self-distillation enable robust integration under incomplete inputs.

2 Methodology

2.1 Overview of KiD-Seg

To address incomplete multi-modal liver MRI, we propose *KiD-Seg*. The full modality set $\mathcal{M}$ is partitioned into four clinically atomic groups: anatomical anchor $\mathcal{G}_1 = \{\texttt{T2WI}\}$, dynamic contrast $\mathcal{G}_2 = \{\texttt{C-pre}, \texttt{C+A}, \texttt{C+V}, \texttt{C+Delay}\}$, diffusion $\mathcal{G}_3 = \{\texttt{DWI}\}$, and Dixon $\mathcal{G}_4 = \{\texttt{InPhase}, \texttt{OutPhase}\}$. Assuming `T2WI` is always acquired, we model group-wise missingness by sampling $\mathcal{S}_G \subseteq \{\mathcal{G}_1, \ldots, \mathcal{G}_4\}$ with $\mathcal{G}_1 \in \mathcal{S}_G$, yielding the available modality subset $\mathcal{S} = \bigcup_{\mathcal{G}_g \in \mathcal{S}_G} \mathcal{G}_g \subseteq \mathcal{M}$. KiD-Seg handles arbitrary $\mathcal{S}$ using three key modules: **AGCL** (anchor-guided anatomical alignment), **DKD** (contrast kinetic modeling), and **HGF** (intra-/inter-group fusion). A tri-branch multi-scale decoder (reconstruction, independent segmentation, fused segmentation) then processes these features, injecting HGF and kinetic cues into the fused branch via gated residual connections.

2.2 Asymmetric Anchor-Guided Contrastive Learning

Clinically, `T2WI` ($\mathcal{G}_1$) provides the most reliable anatomical reference, while auxiliary sequences (e.g., `DWI`) are more distortion-prone. Thus, KiD-Seg treats `T2WI` as a fixed *anatomical anchor*. To preserve anchor fidelity, asymmetric one-way contrastive alignment pulls auxiliary modalities toward the anchor. For an available subset $\mathcal{S}$ (Fig. 1), each modality $m \in \mathcal{S}$ is represented as multi-scale features

$\{F_m^l\}_{l=1}^L$ with a shared encoder. The global anatomical embedding of modality m is derived from the bottleneck feature with global average pooling $\mathrm{Pool}(\cdot)$ and a projection head $g_a(\cdot)$:

$$z_a^m \;=\; g_a\big(\mathrm{Pool}(F_m^L)\big)\,. \tag{1}$$

We optimize a modified InfoNCE loss [13] with asymmetric alignment, pulling each auxiliary modality ($m \in \mathcal{S} \setminus \mathcal{G}_1$) toward the anatomical anchor while preserving the anchor via stop-gradient $\mathrm{sg}(\cdot)$:

$$\mathcal{L}_{\mathrm{AG}} \;=\; \sum_{m\in\mathcal{S}\setminus\mathcal{G}_1} -\log \frac{\exp(\mathrm{sim}(z_a^m,\, \mathrm{sg}(z_a^{\texttt{T2WI}}))/\tau_{\mathrm{AG}})}{\sum_{j=1}^{B} \exp\big(\mathrm{sim}(z_a^m,\, \mathrm{sg}(z_{a,j}^{\texttt{T2WI}}))/\tau_{\mathrm{AG}}\big)}, \tag{2}$$

where B is the batch size, τ_{AG} is the temperature, and $\mathrm{sim}(\boldsymbol{u},\boldsymbol{v}) = \frac{\boldsymbol{u}^\top \boldsymbol{v}}{\|\boldsymbol{u}\|\,\|\boldsymbol{v}\|}$ denotes cosine similarity. This asymmetric loss maps heterogeneous auxiliary modalities into the shared `T2WI`-anchored anatomical space while preserving the anchor representation.

2.3 Differential Kinetic Disentanglement for Dynamic Contrast

To capture tumor hemodynamics (e.g., enhancement/washout) while reducing intensity-scale sensitivity, deformably registered DCE phases are differenced voxel-wise relative to the common `C-pre` reference. Let I^q be the volume at phase $q \in \{\texttt{C+A}, \texttt{C+V}, \texttt{C+Delay}\}$ and $I^{\texttt{C-pre}}$ the pre-contrast volume. The phase-wise difference maps are:

$$\Delta I^q \;=\; I^q - I^{\texttt{C-pre}}. \tag{3}$$

A lightweight 3D difference encoder E_Δ encodes ΔI^q, and cross-phase attention pooling Φ_{kin} aggregates phases into a kinetic feature map z_k:

$$z_k \;=\; \Phi_{\mathrm{kin}}(\{E_\Delta(\Delta I^q)\}_q)\,. \tag{4}$$

The kinetic feature z_k is projected to each decoder scale and injected into the fused decoder via gated residual connections (Sec. 2.5), enabling progressive hemodynamic refinement with improved training stability. To enforce discriminative kinetics, we further apply a **Kinetic Contrastive Loss** $\mathcal{L}_{\mathrm{KCL}}$, computed only when $\mathcal{G}_2$ is available. Specifically, we sample stratified voxels $\mathcal{I} \subset \Omega$ and project voxel-wise kinetic features into a normalized latent space using $g_k(\cdot)$ as $\boldsymbol{z}_i = \frac{g_k(z_k(\mathbf{x}_i))}{\|g_k(z_k(\mathbf{x}_i))\|}$, $i \in \mathcal{I}$, where $\mathbf{x}_i$ is the coordinate of voxel i, Ω is the spatial domain, and y_i denotes the voxel label (from ground truth). For an anchor voxel $i \in \mathcal{I}$, the positive set is defined as $\mathcal{P}(i) = \{p \in \mathcal{I} \setminus \{i\} \mid y_p = y_i\}$. We use a supervised contrastive objective:

$$\mathcal{L}_{\mathrm{KCL}} \;=\; \sum_{i\in\mathcal{I}} \frac{-1}{|\mathcal{P}(i)|} \sum_{p\in\mathcal{P}(i)} \log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_p)/\tau_{\mathrm{K}})}{\sum_{a\in\mathcal{I}\setminus\{i\}} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_a)/\tau_{\mathrm{K}})}, \tag{5}$$

where τ_{K} is the temperature and g_k projects voxel-wise kinetic features into ℓ_2-normalized embeddings, clustering same-class voxels while separating tumor and background kinetic patterns.

2.4 Hierarchical Atomic-Group Fusion

Liver MRI reading is hierarchical: sequence-specific cues are first extracted and then integrated on the T2WI structural scaffold. To emulate this under group-wise missingness, we adopt a two-stage fusion at each scale $l \in \{1, \dots, L\}$: **(i)Intra-Group Fusion (evidence consolidation):** For each group $g \in \{1, \dots, 4\}$, let $\mathcal{M}_g = \mathcal{G}_g \cap \mathcal{S}$. For multi-modality groups (e.g., $\mathcal{G}_2, \mathcal{G}_4$), overlapping yet complementary features are merged by attention-weighted pooling. Let $\Phi^l_{\text{att}}(\cdot)$ be a lightweight network producing voxel-wise importance maps, with weights normalized across modalities in each group:

$$\alpha^l_m(\mathbf{x}) \;=\; \frac{\exp(\Phi^l_{\text{att}}(F^l_m)(\mathbf{x}))}{\sum_{k\in\mathcal{M}_g}\exp(\Phi^l_{\text{att}}(F^l_k)(\mathbf{x}))}, \qquad H^l_g(\mathbf{x}) \;=\; \sum_{m\in\mathcal{M}_g} \alpha^l_m(\mathbf{x})\, F^l_m(\mathbf{x}), \quad (6)$$

where $\mathbf{x} \in \Omega_l$ is the voxel index at scale l; for single-modality groups, this reduces to identity. **(ii) Inter-Group Fusion (Anatomical Querying):** We then fuse consolidated group evidence with the anatomical anchor via masked cross-attention, where T2WI actively queries auxiliary evidence (instead of symmetric concatenation), preserving its role as the primary structural reference. Specifically, $F^l_{\texttt{T2WI}}$ is used as the query, and the auxiliary token set is defined as $\mathcal{T}^l = \{F^l_{\texttt{T2WI}}\} \cup \{H^l_g\}_{g:\,\mathcal{M}_g\neq\varnothing}$. Let $r_g \in \{0, 1\}$ indicate whether group g is present (i.e., $r_g = 1 \iff \mathcal{M}_g \neq \varnothing$); these indicators define an attention mask. The fused feature at scale l is:

$$\widetilde{F}^l \;=\; F^l_{\texttt{T2WI}} \;+\; \text{CrossAttn}\Big(Q = F^l_{\texttt{T2WI}},\; K = V = \mathcal{T}^l,\; \text{mask} = \mathbf{r}\Big), \quad (7)$$

where $\mathbf{r} = (r_1, \dots, r_4)$. This design lets the anatomical anchor query auxiliary cues only when available, reducing noise propagation under missing modalities.

2.5 Tri-Branch Multi-Scale Decoder with Gated HGF and DKD

Using Eq. (7) only at the bottleneck may miss fine boundaries; thus, we extend HGF to a multi-scale decoder and inject fused anatomical evidence and kinetic cues at all scales. **(i) Scale-wise Kinetic Projection:** The kinetic feature z_k (Eq. (4)) is projected to each decoder scale as $K^l = \psi^l_k(\text{Up}_l(z_k))$, $l \in \{1, \dots, L\}$, where $\text{Up}_l(\cdot)$ is scale-specific upsampling and $\psi^l_k(\cdot)$ is a learnable 1×1×1 projection; if $\mathcal{G}_2$ is unavailable, $K^l = \mathbf{0}$. **(ii) Three Decoder Branches:** Following [8], the decoder has three branches: independent segmentation from the base stream, fused segmentation with HGF and DKD cues, and reconstruction. Let D^l_{base} be the base decoder feature at scale l from a standard top-down decoder with skip connections. The segmentation branch is defined as $D^l_{\text{ind}} = D^l_{\text{base}}$. To improve training stability under frequent missingness, we inject HGF and DKD via *gated residual connections*:

$$D^l_{\text{fus}} \;=\; D^l_{\text{base}} \;+\; \sigma(\gamma_l)\cdot\widetilde{F}^l \;+\; \sigma(\beta_l)\cdot K^l, \quad (8)$$

where $\widetilde{F}^l$ is the HGF output from Eq. (7), $\sigma(\cdot)$ is the sigmoid function, and γ_l, β_l are learnable scalar gates. The reconstruction branch uses modality-specific

heads to recover the observed inputs, i.e., $\hat{I}_m = h_{\mathrm{rec},m}\big(D^1_{\mathrm{base}}\big)$ for $m \in \mathcal{S}$, where $h_{\mathrm{rec},m}$ denotes the reconstruction head for modality m. **(iii)Prediction Heads:** At each decoder scale, we attach two segmentation heads, namely $\hat{P}^{\mathrm{ind}}_l = h^l_{\mathrm{ind}}\big(D^l_{\mathrm{ind}}\big)$ and $\hat{P}^{\mathrm{fus}}_l = h^l_{\mathrm{fus}}\big(D^l_{\mathrm{fus}}\big)$, with corresponding probabilistic outputs $\hat{Y}^{\mathrm{ind}}_l = \mathrm{softmax}(\hat{P}^{\mathrm{ind}}_l)$ and $\hat{Y}^{\mathrm{fus}}_l = \mathrm{softmax}(\hat{P}^{\mathrm{fus}}_l)$. The fused branch supports evaluation and self-distillation, while the training-only independent branch prevents reliance on potentially unavailable HGF/DKD cues.

2.6 Completeness-Aware Self-Distillation and Objective

To reduce the gap between incomplete and complete inference, we use self-distillation with structured group dropout: a student pass samples missing groups to form $\mathcal{S}$, while a parallel teacher pass uses a maximal set $\mathcal{S}^{\max} \subseteq \mathcal{M}$ (typically $\mathcal{M}$). Both predictions are taken from the final fused branch, $\hat{P}^{\max} \equiv \hat{P}^{\mathrm{fus},\max}_1$ and $\hat{P}^{\mathcal{S}} \equiv \hat{P}^{\mathrm{fus},\mathcal{S}}_1$, and optimized via voxel-wise KL divergence:

$$\mathcal{L}_{\mathrm{CA}} \;=\; T^2 \cdot \frac{1}{B|\Omega|} \sum_{b=1}^{B} \sum_{v\in\Omega} \mathrm{KL}\Big(\mathrm{softmax}(\mathrm{sg}(\hat{P}^{\max}_{b,:,v})/T) \;\Big\|\; \mathrm{softmax}(\hat{P}^{\mathcal{S}}_{b,:,v}/T)\Big)\,, \tag{9}$$

where T is the temperature, $\mathrm{sg}(\cdot)$ is the stop-gradient operator, B is batch size, and Ω is the spatial voxel set. We jointly optimize the reconstruction, independent-segmentation, and fused-segmentation branches with branch-specific objectives. Let $\hat{I}_m$ and I_m denote the reconstructed and input images for modality $m \in \mathcal{S}$, respectively. **The reconstruction branch** is supervised by an ℓ_1 loss over observed modalities, i.e., $\mathcal{L}_{\mathrm{rec}} = \frac{1}{|\mathcal{S}|\,|\Omega|} \sum_{m\in\mathcal{S}} \sum_{v\in\Omega} \Big|\hat{I}_m(v) - I_m(v)\Big|$. This branch regularizes shared decoder features to preserve modality-specific appearance under incomplete inputs. Since the independent and fused segmentation branches share the same deep-supervision loss structure, we unify them using a branch index $b \in \{\mathrm{ind}, \mathrm{fus}\}$ as $\mathcal{L}^b_{\mathrm{seg}} = \sum_{l=1}^{L} \Big(\mathcal{L}_{\mathrm{Dice}}(\hat{Y}^b_l, Y) + \mathcal{L}_{\mathrm{WCE}}(\hat{Y}^b_l, Y)\Big)\,,\ b \in \{\mathrm{ind}, \mathrm{fus}\}$, where $\mathcal{L}_{\mathrm{WCE}}$ is a class-weighted cross-entropy term used to mitigate lesion–background imbalance, $b = \mathrm{ind}$ denotes the **independent segmentation branch** (without HGF/DKD), and $b = \mathrm{fus}$ denotes the **fused segmentation branch** (with HGF and DKD gated residual injection). The fused branch is the main prediction branch used for evaluation and for completeness-aware self-distillation. The **final objective** combines branch-specific supervision and regularization using unit-weighted loss terms:

$$\mathcal{L}_{\mathrm{total}} \;=\; \mathcal{L}^{\mathrm{fus}}_{\mathrm{seg}} \;+\; \mathcal{L}^{\mathrm{ind}}_{\mathrm{seg}} \;+\; \mathcal{L}_{\mathrm{rec}} \;+\; \mathcal{L}_{\mathrm{AG}} \;+\; \mathcal{L}_{\mathrm{KCL}} \;+\; \mathcal{L}_{\mathrm{CA}}. \tag{10}$$

3 Experiments

3.1 Experimental Setup

We evaluate KiD-Seg on the LLD-MMRI dataset [11], a 498-case multi-parametric liver MRI cohort containing seven lesion categories: hepatocellular carcinoma,

Table 1. Ablation study reporting overall results across all valid group-wise modality combinations. † indicates that the metric-wise best method significantly outperforms the corresponding variant ($p < 0.05$). The best results (by median) are in **bold**.

Method Variant	DSC ↑	HD_{95} (mm) ↓
Baseline (DC-Seg [8])	0.648 ± 0.304 (0.764)†	11.559 ± 21.085 (5.000)†
+ AGCL	0.677 ± 0.290 (0.769)†	10.447 ± 20.188 (4.899)†
+ AGCL + DKD	0.681 ± 0.278 (0.773)†	15.200 ± 28.369 (5.099)†
+ AGCL + DKD + HGF (Full KiD-Seg)	$\mathbf{0.670 \pm 0.291}$ (**0.781**)	$\mathbf{8.737 \pm 16.189}$ (**4.123**)

Table 2. Quantitative comparison with SOTA methods. † indicates that the row-wise best method significantly outperforms the corresponding method ($p < 0.05$). Best and second-best median results are shown in **bold** and underlined, respectively. •/◦ denote available/missing modalities, and parentheses report parameter counts (M).

Atomic Groups				DSC ↑				
$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	RFNet(14.91M) [6]	mmFormer(250.18M) [18]	M^3AE(2.20M) [9]	DC-Seg(76.79M) [8]	KiD-Seg(77.02M)
•	◦	◦	◦	**0.561 ± 0.367** (**0.737**)	0.557 ± 0.363 (0.709)†	0.428 ± 0.353 (0.530)†	0.558 ± 0.359 (0.733)†	0.600 ± 0.326 (0.730)†
•	◦	◦	•	0.627 ± 0.333 (0.756)†	**0.652 ± 0.316** (**0.796**)	0.521 ± 0.367 (0.654)†	0.653 ± 0.299 (0.757)†	0.645 ± 0.309 (0.752)†
•	◦	•	◦	0.634 ± 0.331 (0.772)†	0.654 ± 0.305 (0.770)†	0.513 ± 0.362 (0.696)†	0.677 ± 0.279 (0.770)†	**0.664 ± 0.284** (**0.776**)
•	•	◦	◦	0.659 ± 0.316 (0.772)†	0.661 ± 0.291 (0.764)†	0.537 ± 0.363 (0.682)†	0.638 ± 0.307 (0.766)†	**0.678 ± 0.291** (**0.773**)
•	◦	•	•	0.661 ± 0.314 (0.768)†	**0.670 ± 0.295** (**0.793**)	0.551 ± 0.358 (0.727)†	0.681 ± 0.279 (0.759)†	0.701 ± 0.262 (0.785)†
•	•	◦	•	0.666 ± 0.312 (0.774)†	**0.674 ± 0.290** (**0.781**)	0.554 ± 0.363 (0.714)†	0.650 ± 0.302 (0.763)†	0.667 ± 0.299 (0.778)†
•	•	•	◦	0.702 ± 0.280 (0.772)†	0.689 ± 0.276 (0.781)†	0.597 ± 0.342 (0.770)†	0.658 ± 0.296 (0.772)†	**0.703 ± 0.268** (**0.790**)
•	•	•	•	0.702 ± 0.276 (0.778)†	0.689 ± 0.281 (0.779)†	0.615 ± 0.334 (0.761)†	0.669 ± 0.287 (0.768)†	**0.702 ± 0.265** (**0.802**)
Overall				0.651 ± 0.320 (0.771)†	0.656 ± 0.306 (0.773)†	0.540 ± 0.359 (0.694)†	0.648 ± 0.304 (0.764)†	**0.670 ± 0.291** (**0.781**)

intrahepatic cholangiocarcinoma, liver metastasis, hepatic hemangioma, hepatic abscess, hepatic cyst, and focal nodular hyperplasia. Together with background, these constitute eight segmentation labels and include both malignant and benign lesions. We use a patient-level 7:2:1 train/validation/test split and report test-set results. Volumes are resampled to $256 \times 256 \times 32$ ($1.4 \times 1.4 \times 1.7$ mm), SyN-registered with annotations to the common `C-pre` space [1], and min–max normalized per modality; `T2WI` serves only as the representation-learning anchor. To mimic clinical incompleteness, we enforce group-wise modality availability with atomic blocks $\mathcal{G}_1$={`T2WI`} (always present), $\mathcal{G}_2$ (dynamic contrast), $\mathcal{G}_3$={`DWI`}, and $\mathcal{G}_4$ (Dixon), and sample valid combinations with $\mathcal{G}_1$ always included. KiD-Seg is trained on one NVIDIA A100 GPU (80 GB) using AdamW (lr 2×10^{-4}, wd 1×10^{-5}), cosine annealing, 200 epochs, and batch size 2. The shared encoder depth is set to 4 (i.e., L=4). At inference, the same group-aware fusion handles any valid modality block set (including `T2WI`-only) without retraining. We evaluate segmentation with Dice similarity coefficient (DSC) and the 95^{th} percentile Hausdorff distance (HD_{95}, mm), reported as mean±std (median) over test subjects. Statistical significance is assessed by a two-sided Wilcoxon signed-rank test ($p < 0.05$).

3.2 Ablation Studies

Ablation results in Table 1 show metric-dependent effects of the proposed components. AGCL improves mean/median DSC from 0.648/0.764 to 0.677/0.769,

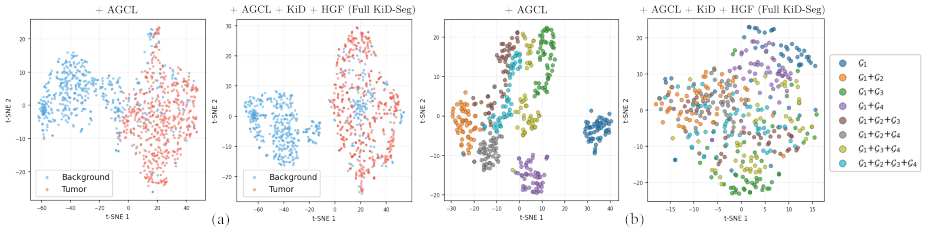

Fig. 2. t-SNE feature analysis. (a) Class-wise embeddings under `T2WI`-only input, showing improved tumor–background separation with full KiD-Seg. (b) Tumor embeddings colored by modality combination, where greater overlap indicates stronger cross-combination invariance.

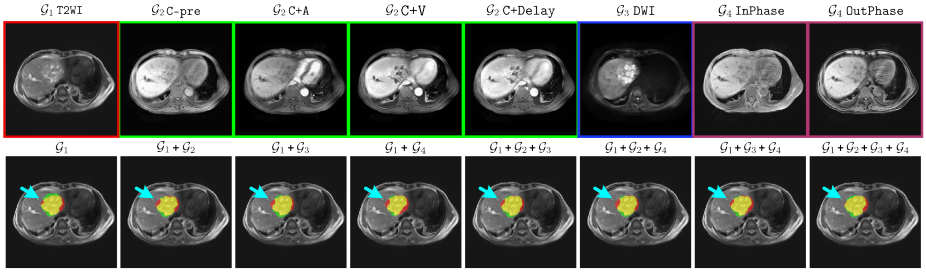

Fig. 3. KiD-Seg qualitative results. Top: 8-modal liver MRI. **Bottom**: T2WI overlays for all valid combinations (green: prediction, red: ground truth, yellow: overlap), showing stable tumor segmentation under group-wise missingness (blue arrow).

indicating the benefit of anchor-guided alignment. Adding DKD further improves DSC to 0.681/0.773, although HD_{95} increases to 15.200/5.099 mm. With HGF, the full KiD-Seg model achieves the best median DSC of 0.781 and the best HD_{95} of 8.737/4.123 mm, despite a slight decrease in mean DSC to 0.670. These results suggest that HGF improves typical-case consistency and boundary robustness, while Fig. 2 shows enhanced separability and feature invariance.

3.3 Comparison with State-of-the-Art

Quantitative results (Table 2) show that KiD-Seg achieves the best overall median DSC (0.781) and the best full-modality median DSC (0.802), while remaining competitive across group-wise missingness settings. Although mmFormer leads on several specific combinations, KiD-Seg achieves the best overall and full-modality median DSC, with particularly strong results for $\mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3$ (0.790) and full-modality inputs (0.802). Under identical resolution and hardware, KiD-Seg adds only 0.23M parameters over DC-Seg while improving the overall median DSC from 0.764 to 0.781. Fig. 3 demonstrates robust localization under $\mathcal{G}_1$-only input and progressive boundary refinement with added auxiliary groups.

4 Conclusion

We proposed KiD-Seg, a clinically grounded framework for liver tumor segmentation under group-wise MRI missingness. By anchoring anatomical learning to `T2WI` and modeling auxiliary sequences as atomic groups, AGCL, DKD, and HGF improve alignment, tumor discrimination, and cross-combination robustness. On the 498-case LLD-MMRI dataset, KiD-Seg achieves the best overall and full-modality median DSCs of 0.781 and 0.802, respectively, with only 0.23M more parameters than DC-Seg. Limitations include reliance on a fixed `T2WI` anchor, sensitivity to inter-phase registration errors, and single-cohort evaluation; future work will investigate anchor-flexible, registration-robust learning and multi-center validation.

Acknowledgments. This work was supported by Ant Group, the Ant Group Research Intern Program, and Zhongshan Hospital, Fudan University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical image analysis **12**(1), 26–41 (2008)
2. Azad, R., Khosravi, N., Dehghanmanshadi, M., Cohen-Adad, J., Merhof, D.: Medical image segmentation on mri images with missing modalities: A review. arXiv preprint arXiv:2203.06217 (2022)
3. Azad, R., Khosravi, N., Merhof, D.: Smu-net: Style matching u-net for brain tumor segmentation with missing modalities. In: International conference on medical imaging with deep learning. pp. 48–62. PMLR (2022)
4. Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.A.: Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: International conference on medical image computing and computer-assisted intervention. pp. 447–456. Springer (2019)
5. Chernyak, V., Fowler, K.J., Kamaya, A., Kielar, A.Z., Elsayes, K.M., Bashir, M.R., Kono, Y., Do, R.K., Mitchell, D.G., Singal, A.G., et al.: Liver imaging reporting and data system (li-rads) version 2018: imaging of hepatocellular carcinoma in at-risk patients. Radiology **289**(3), 816–830 (2018)
6. Ding, Y., Yu, X., Yang, Y.: Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3975–3984 (2021)
7. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: Hemis: Hetero-modal image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 469–477. Springer (2016)
8. Li, H., Li, Z., Mao, Y., Ding, Z., Huang, Z.: Dc-seg: Disentangled contrastive learning for brain tumor segmentation with missing modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 138–148. Springer (2025)

9. Liu, H., Wei, D., Lu, D., Sun, J., Wang, L., Zheng, Y.: M3ae: multimodal representation learning for brain tumor segmentation with missing modalities. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 1657–1665 (2023)
10. Liver, E.A.F.T.S.O.T., et al.: Easl clinical practice guidelines on the management of hepatocellular carcinoma. Journal of hepatology **82**(2), 315–374 (2025)
11. Lou, M., Ying, H., Liu, X., Zhou, H.Y., Zhang, Y., Yu, Y.: Sdr-former: A siamese dual-resolution transformer for liver lesion classification using 3d multi-phase imaging. Neural Networks p. 107228 (2025)
12. Meng, X., Sun, K., Xu, J., He, X., Shen, D.: Multi-modal modality-masked diffusion network for brain mri synthesis with random modality missing. IEEE Transactions on Medical Imaging **43**(7), 2587–2598 (2024)
13. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
14. Shen, L., Zhu, W., Wang, X., Xing, L., Pauly, J.M., Turkbey, B., Harmon, S.A., Sanford, T.H., Mehralivand, S., Choyke, P.L., et al.: Multi-domain image completion for random missing input data. IEEE transactions on medical imaging **40**(4), 1113–1122 (2020)
15. Taouli, B., Koh, D.M.: Diffusion-weighted mr imaging of the liver. Radiology **254**(1), 47–66 (2010)
16. Van Tulder, G., de Bruijne, M.: Why does synthesized data improve multi-sequence classification? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 531–538. Springer (2015)
17. Yang, Q., Guo, X., Chen, Z., Woo, P.Y., Yuan, Y.: D 2-net: Dual disentanglement network for brain tumor segmentation with missing modalities. IEEE Transactions on Medical Imaging **41**(10), 2953–2964 (2022)
18. Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., Zheng, Y.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 107–117. Springer (2022)