OT-MIMIC: LEARNING MEDICAL IMAGE SEGMENTATION WITH MISSING MODALITIES VIA STRUCTURE-AWARE OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal vision–language models (VLMs) for medical image segmentation often assume fully paired image–text data, an assumption rarely satisfied in real-world clinical datasets. In practice, large-scale medical corpora contain abundant partially observed samples and only a small subset of fully paired records. We study medical image segmentation under this incomplete multimodal setting and propose OT-MIMIC, a structure-aware transport-based completion framework that learns from partially observed data rather than discarding it. OT-MIMIC maps incomplete records to a dynamically refined memory bank and reconstructs missing representations or supervision through barycentric aggregation. We instantiate the framework with an asymmetric Fused Gromov–Wasserstein (FGW) objective in a CLIP-aligned latent space, combining semantic correspondence with relational consistency between observed-side and complementary-side structure. We additionally derive a high-probability excess-risk bound that decomposes learning error into empirical risk minimization, transport completion, recoverability, and function-approximation terms. Experiments on QaTa-COV19 and MosMedData+ show strong performance under both missing-text and missing-image-and-mask settings as complete-pair availability decreases from 50% to 1%, with the largest gains under extreme scarcity. Controlled ablations further show that the FGW instantiation improves over nearest-anchor retrieval and feature-mean completion under an otherwise fixed training pipeline.

1 INTRODUCTION

Modern data is characterized by multiple modalities that provide rich, diverse information. Deep learning advancements, such as Vision-Language Models (VLMs), have empowered AI to harness this complementary information, achieving performance that far exceeds that of individual modalities. However, the efficacy of these multimodal methods hinges on the availability of vast quantities of high-quality data with well-aligned cross-modal features Zhang et al. (2023); Ma et al. (2024). Unfortunately, real-world datasets often deviate from this ideal, and missing modalities are widespread.

These challenges are particularly significant in the medical domain due to stringent constraints in data acquisition and curation Qian et al. (2025). For instance, obtaining a Positron Emission Tomography (PET) scan involves substantial costs and requires specialized expert annotation. In contrast, Electronic Health Records (EHRs) are routinely generated during patient care, making them relatively abundant. Furthermore, while clinical reports describing the pathology of images have proven highly valuable, they are frequently missing from public cohorts due to various logistical and privacy limitations.

The aforementioned intertwined difficulties pose significant challenges to existing methodologies. Simply discarding samples with missing modalities is inefficient and wasteful. While large-scale self-supervised Vision-Language Pretraining (VLP) models can leverage multi-modal datasets by aligning features via contrastive loss Wang et al. (2022); Zhang et al. (2023), these frameworks typically assume the availability of **full modalities**. Consequently, they fail to fully exploit data where specific modalities are absent.

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.

Classical statistical methods are supported by rigorous theoretical foundations Little & Rubin (2019), yet they often rely on strict parametric or semi-parametric priors. These assumptions prevent them from scaling to the extremely high-dimensional, heterogeneous data characteristic of the modern deep learning era. Recent deep learning approaches attempt to mitigate these issues by either learning modality-agnostic features through domain adversarial losses Wang et al. (2021); Konwer et al. (2023), or by explicitly generating the missing modality from the available ones Ma et al. (2021); Wang et al. (2023b); Kim & Kim (2024). However, these methods primarily focus on synthesizing information at the **individual datum level**, which we argue is insufficient. The fundamental principle of multi-modal learning is that distinct modalities provide complementary information. Therefore, predicting a missing modality from a single observed sample is generally underdetermined and can suppress modality-specific information that is better captured at the distribution level.

In this paper, we address the missing-modality problem through transport-based completion. Rather than discarding incomplete records or predicting each missing component from a single observed sample, OT-MIMIC maps partially observed records toward a memory bank that approximates the complete-data distribution. The resulting transport weights are then used for barycentric aggregation, allowing the same completion mechanism to recover missing representations and missing segmentation supervision.

The framework separates the general completion strategy from the particular transport geometry. In our main implementation, we use Fused Gromov–Wasserstein (FGW) because it combines feature compatibility with relational structure. Standard feature-based transport compares individual samples, whereas FGW additionally encourages pairs that are similar on the observed side to remain structurally compatible on the complementary side. We therefore define an asymmetric FGW objective whose source relations are computed from observed modalities and whose target relations are computed from the modalities to be completed. Each missing component is recovered by the transport-weighted barycentric projection over the memory bank.

Building on this transport-completion view, we develop a multimodal learning framework for severe modality imbalance. We validate it in multimodal segmentation and extend the same mechanism to missing supervision by barycentrically aggregating memory-bank masks with the learned transport weights.

Our key contributions are summarized as follows:

- We introduce OT-MIMIC, a transport-based completion framework that converts partially observed vision–language records into usable training examples through memory-bank barycentric aggregation.
- We instantiate the framework with an asymmetric FGW objective that combines observed-modality feature compatibility with complementary-side relational structure, providing a structure-aware alternative to purely feature-based completion.
- We provide a finite-sample excess-risk analysis that separates ERM concentration, transport-completion, recoverability, and function-approximation errors.
- We combine transport-based completion with a self-refining uncertainty-curated memory bank and show robust performance under extreme missing-text and missing-image-and-mask regimes. Controlled and cross-domain extensions are reported in the Appendix.

2 Related Work

Multi-modal Medical Image Segmentation and Pretraining. Text-guided segmentation couples visual features with language grounding Hu et al. (2016); Yang et al. (2022b). In medicine, LViT Li et al. (2023) uses text-assisted pseudo supervision, DuSSS Pan et al. (2025) develops VLM-guided semi-supervised segmentation, and MedCLIPSeg Koleilat et al. (2026) adapts CLIP for data-efficient dense prediction. Foundation-model lines include MedSAM Ma et al. (2024), MedCLIP-style pretraining Wang et al. (2022); Zhang et al. (2023), and 3D vision–language pretraining Blankemeier et al. (2024). These works motivate strong multimodal representations, but do not directly address the group-wise image/text incompleteness together with missing segmentation supervision studied here.

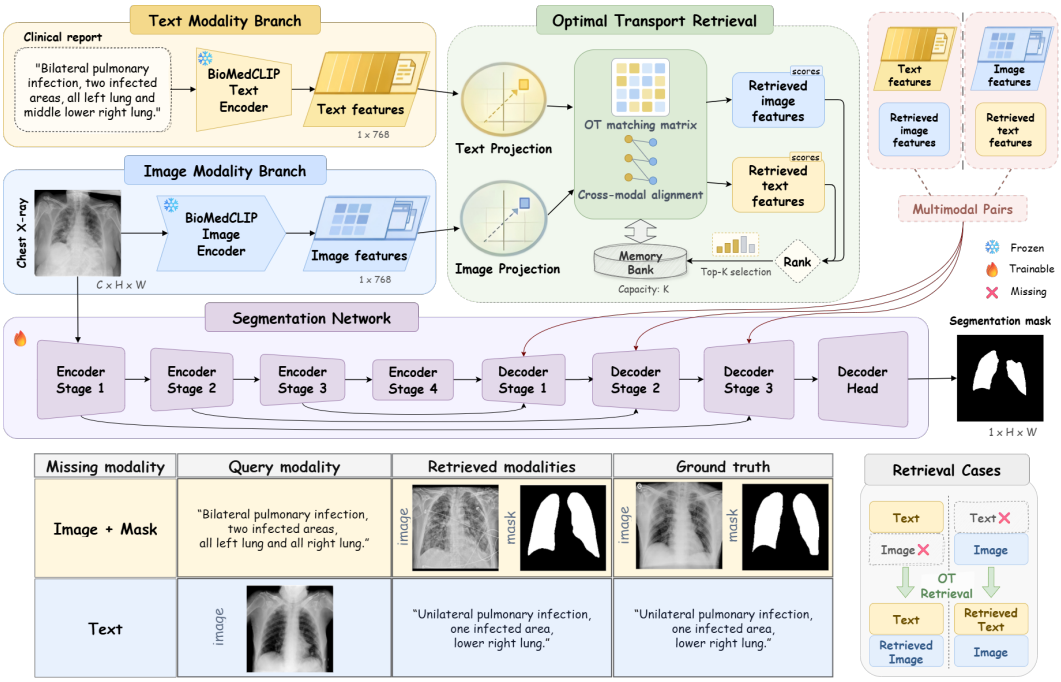

Figure 1: **OT-MIMIC architecture overview.** BiomedCLIP encoders extract image and text features. An FGW module with a memory bank computes a soft assignment $\widehat{\mathbf{T}}$ to all bank anchors and completes each missing modality by the barycentric projection $\hat{\mathbf{x}} = \sum_k w_k^* \mathbf{z}_k$. *Rank* sorts bank entries by predictive uncertainty for admission and eviction under capacity K.

Learning with Missing Modalities. Missing-modality learning has been widely studied in medical image analysis, especially for incomplete multi-sequence MRI segmentation. Existing approaches include hetero-modal fusion Havaei et al. (2016), disentangled representation learning Yang et al. (2022a), modality-adaptive interaction Zhao et al. (2022), Transformer-based fusion Zhang et al. (2022), prototype distillation Wang et al. (2023a), and masked auto encoding/self-distillation Liu et al. (2023). More recent work further considers clinically realistic imbalanced missing rates across modalities Shi et al. (2024). However, these methods are largely restricted to image-only modality sets. In the vision-language setting, DiCMoR Wang et al. (2023b) imposes a Gaussian terminal prior, whereas our transport formulation does not require this parametric distributional assumption. ProLearn Ye et al. (2025) alleviates missing text supervision by retrieving semantically related text cues under a prototype learning framework. However, it mainly relies on feature-level similarity and does not explicitly model global cross-modality neighborhood structure. Moreover, existing methods do not address the more general setting where images, text, and segmentation supervision can be simultaneously missing in a group-wise manner. Although recent multimodal representation learning has started to jointly consider missing modalities and missing labels Wu et al. (2024), these methods focus on patient-level prediction rather than dense segmentation. In contrast, we study vision–language medical image segmentation under joint modality and supervision incompleteness within a unified transport-completion and learning framework.

Machine Learning with Optimal Transport. Optimal Transport (OT) provides a principled framework for comparing probability measures by searching for couplings that "move mass" from one distribution to another at minimum cost Villani et al. (2008). The induced p-Wasserstein distance serves as a fundamental metric for probability measures. To alleviate the computational cost of exact OT, entropic regularization Cuturi (2013) enables efficient optimization on modern hardware via the Sinkhorn algorithm. Gromov-Wasserstein (GW) problem Mémoli (2011) allows transport between distributions residing on disparate metric spaces, Fused Gromov-Wasserstein (FGW) Vayer et al. (2020) further extends GW by blending a conventional feature-level transport cost with the structural GW distortion term. This dual alignment makes FGW well suited to cross-modal completion when feature compatibility and relational geometry are both informative.

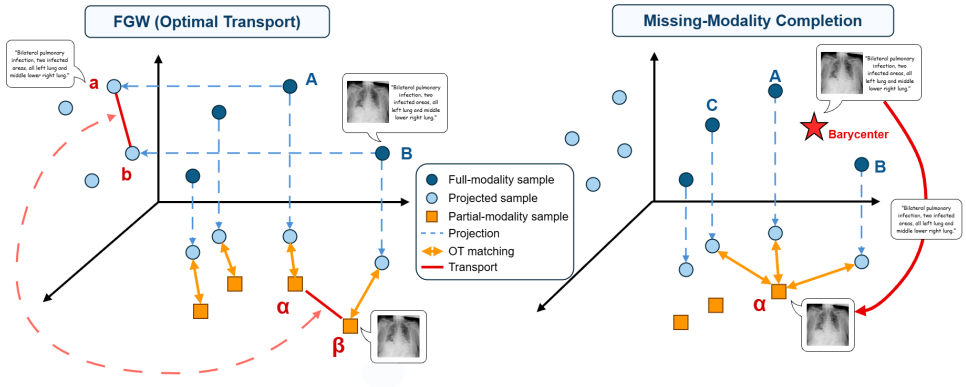

Figure 2: **Optimal transport for missing-modality completion. Left:** illustration of FGW. **Right: Missing Modality Completion:** Since the OT framework does not enforce one-to-one matching, partial-modality data α may be matched to multiple full-modality data (A, B and C). Their barycenter is calculated (red star) to complete the missing modality for α.

3 Methodology

3.1 Optimal Transport and Fused Gromov-Wasserstein Distance

Optimal Transport (OT) provides a principled framework for measuring discrepancies between probability distributions and constructing sample-level correspondences via an optimal coupling. Classical Wasserstein-based OT requires both distributions to live on a common metric space, which is rarely true across modalities. We therefore adopt the *Fused Gromov-Wasserstein* (FGW) distance Vayer et al. (2020), which combines a feature-level transport cost with a structure-preserving Gromov-Wasserstein distortion. The feature term captures semantic compatibility while the structural term preserves relational geometry. This is useful for cross-modal completion when feature similarity alone is ambiguous because FGW accounts for both sample-level similarity and pairwise structure.

Figure 2 illustrates this idea. Complete bank records provide paired representations across the observed and complementary spaces, while partially observed queries are available only in the observed space. Feature-level OT favors semantic compatibility between a query and bank anchors, while the GW term compares pairwise geometry across the observed and complementary spaces. For example, if complete records A and B are matched to partially observed records α and β, the structural term compares the corresponding pairwise distances in the two spaces. The fused objective therefore balances sample-level similarity with relational consistency.

Concretely, let $\mathbf{X}^s \subset \mathbb{R}^{d_s}$ and $\mathbf{X}^t \subset \mathbb{R}^{d_t}$ be source/target sample sets with marginals $\mathbf{p}, \mathbf{q}$, cross-domain ground cost $\mathbf{M}_{s,t}$, and intra-domain pairwise distance matrices $\mathbf{C}^s, \mathbf{C}^t$ (the structure matrices of Vayer et al. (2020)). For $\alpha \in [0,1]$, the discrete FGW problem is

$$\begin{aligned} \min_{\mathbf{T}\in\Pi(\mathbf{p},\mathbf{q})} \mathcal{L}_{\mathrm{FGW}}(\mathbf{T}) =& (1-\alpha)\langle \mathbf{T}, \mathbf{M}_{s,t}\rangle_F + \alpha \sum_{i,j,k,l} |\mathbf{C}^s_{i,k} - \mathbf{C}^t_{j,l}|^2 \mathbf{T}_{i,j}\mathbf{T}_{k,l} \\ \text{s.t. } \Pi(\mathbf{p},\mathbf{q}) =& \{\mathbf{T} \in \mathbb{R}_+^{n_s \times n_t} : \mathbf{T}\mathbf{1}_{n_t} = \mathbf{p}, \mathbf{T}^\top \mathbf{1}_{n_s} = \mathbf{q}\} \end{aligned} \tag{1}$$

where $\mathbf{T}$ is the transport plan and $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. The objective is non-convex in $\mathbf{T}$ due to the GW quadratic term. We use the entropic-regularized version $\mathcal{L}_{\mathrm{FGW}}(\mathbf{T}) + \epsilon H(\mathbf{T})$ with $H(\mathbf{T}) = \sum_{i,j} \mathbf{T}_{i,j}(\log \mathbf{T}_{i,j} - 1)$ and $\epsilon = 0.1$, solved by the standard alternating scheme that linearizes the GW term and runs a Sinkhorn inner step on the resulting OT subproblem Vayer et al. (2020). The empirical solver returns a stationary point of this regularized objective, which we denote $\widehat{\mathbf{T}}$ throughout the paper. It is not necessarily a global FGW optimum. We reserve $T^\star$ for the population-level reference completion map used in the proof of Theorem 2 in Appendix B.

3.2 Transport-Based Modality Completion

We address medical image foreground segmentation with Vision–Language Models (VLMs) under incomplete multimodal data and present OT-MIMIC as a *transport-based completion framework*. The

core idea is to map partially observed records to a memory bank that represents the complete-data distribution, then reuse the resulting transport weights to reconstruct missing representations or supervision before standard segmentation training. This framework is conceptually independent of a single transport geometry. Our empirical implementation uses asymmetric FGW because it can exploit both feature compatibility and cross-modal relational structure. The finite-sample analysis is likewise stated for an abstract bounded-complexity completion map rather than for FGW alone. Our empirical study focuses on a bi-modal image–text setting. The same observed-set and complementary-set construction extends naturally to M modalities. Each record exposes an observed subset $\mathcal{O}_c \subseteq \{x^{\text{img}}, x^{\text{txt}}, y\}$ from one of six patterns: complete, missing-mask, missing-image, missing-text, text-only, image-only. These patterns are obtained by zeroing any subset of the three slots except the empty pattern, for example missing-text gives $\mathcal{O}_c = \{x^{\text{img}}, y\}$:

$$\mathcal{O}_c \in \big\{\{x^{\text{img}}, x^{\text{txt}}, y\},\ \{x^{\text{img}}, x^{\text{txt}}\},\ \{x^{\text{txt}}, y\},\ \{x^{\text{img}}, y\},\ \{x^{\text{txt}}\},\ \{x^{\text{img}}\}\big\}, \tag{2}$$

with $x^{\text{img}} \in \mathbb{R}^{C\times H\times W}$, $x^{\text{txt}} \in \mathbb{Z}_+^L$ a tokenized sequence of length L, and $y \in \{0,1\}^{H\times W}$. The task is semi-supervised multimodal learning under group-wise feature and label missingness. The two primary experimental regimes are the *missing-text* pattern $\mathcal{O}_c = \{x^{\text{img}}, y\}$ and the *missing-image-and-mask* pattern $\mathcal{O}_c = \{x^{\text{txt}}\}$.

Model Architecture. As shown in Figure 1, image/text modality completion happens in a shared CLIP latent space: a frozen BiomedCLIP encoder produces penultimate features for each *observed* modality, and a jointly-trained 2-layer GELU MLP (supervised by an image–text matching objective) lifts them to a common d-dimensional space so that the ground cost $\mathbf{M}$ and pairwise matrices $\mathbf{C}^s, \mathbf{C}^t$ in eq. equation 1 are commensurable. Each memory-bank record also stores a (pseudo-)mask y_k. In our bi-modal implementation, masks do not enter the FGW cost or structure matrices. When mask supervision is missing, the resulting row-normalized transport weights are reused to barycentrically aggregate bank masks. The downstream segmentation network is a ConvNeXt-based Liu et al. (2022) U-Net with language-guided fusion Zhong et al. (2023) that consumes the modality-completed feature tuples.

Memory Bank Design. The memory bank serves as a progressively refined empirical proxy for the complete-data distribution: it is seeded with complete cases, and entries are admitted or evicted according to predictive uncertainty, allowing the bank to track the model's evolving representations across epochs. Formally, the bank is a priority queue of capacity K,

$$\mathcal{Z} = \left\{ \left(\eta_k, \mathbf{z}_k^{\text{img}}, \mathbf{z}_k^{\text{txt}}, y_k, e_k, c_k \right) \right\}_{k=1}^{K}, \tag{3}$$

where η_k is the sample index, $\mathbf{z}_k^{\text{img}}, \mathbf{z}_k^{\text{txt}} \in \mathbb{R}^d$ are projected image/text embeddings, $y_k \in \{0,1\}^{H\times W}$ is the (pseudo-)mask, e_k is a per-sample cross-entropy uncertainty (clipped to $[0, 500]$ for numerical stability), and $c_k \in \{0,1\}$ flags whether the record's image, text, *and* mask are all originally observed (i.e., $c_k = 1$ only for records with no completed component). For the bi-modal experiments, the FGW pairwise distance matrices in eq. equation 5 run over the embedded image and text modalities. When the mask is in $\mathcal{O}^c$, $\mathbf{C}^t$ averages over whichever embedded modalities are also in $\mathcal{O}^c$, and the mask is completed by the same row-normalized barycentric rule $\hat{y}_i = \mathbf{1}\{\sum_k w^*_{i,k} y_k \geq 0.5\}$ on the resulting transport plan. With completeness bonus $\lambda = 1000 > \sup e_k$, the retention score $r_k = e_k - \lambda c_k$ enforces a strict ordering: any originally-complete record is preferred over any record with at least one completed component, regardless of relative uncertainty. Among records with the same c_k, lower e_k wins. Originally-complete records can therefore never be evicted by modality-completed records. Pseudo-mask records can be evicted by other pseudo-mask records of lower uncertainty. The bank is recomputed at the end of each epoch using the current model's predictions on the full training set (no test-set leakage), and stored embeddings/features are recomputed each epoch from the current model rather than cached.

Missing Modality Completion Mechanism. For a record with observed subset $\mathcal{O} \subseteq \{1, \ldots, M\}$ and complementary set $\mathcal{O}^c$, OT-MIMIC completes each missing modality $l \in \mathcal{O}^c$ through a transport map $T : \mathcal{H} \to \mathcal{B}$ from observed-modality to complementary-modality embeddings over a memory bank of anchors. The framework abstracts the completion map from its solver. It can use OT-map distillation with a parametric T_θ, kernel or ridge regression on transport weights, or another bounded-complexity parametric family. Theorem 1 provides excess-risk control under this abstract setting.

Our reported experiments instantiate the framework with empirical FGW barycentric projection over the bank, computed per batch with the POT solver. This implementation is non-parametric and falls outside the theorem, with the gap discussed in Appendix B.6. We run one FGW solve per missingness pattern, so $\mathcal{O}$ is fixed across the batch. Each bank record k stores all-modality embeddings $\{\mathbf{z}_k^{(r)}\}_{r=1}^{M}$ that are ℓ_2-normalized in projected BiomedCLIP space.

The ground cost $\mathbf{M}_{i,k}$ in eq. equation 1 averages same-modality squared-ℓ_2 distance over observed modalities only,

$$\mathbf{M}_{i,k} \;=\; \frac{1}{|\mathcal{O}|}\sum_{j\in\mathcal{O}} \big\|\mathbf{x}_i^{(j)} - \mathbf{z}_k^{(j)}\big\|^2. \tag{4}$$

and the pairwise structure matrices are split asymmetrically—the source uses observed embeddings (the only thing it has), the target uses the complementary embeddings to be completed:

$$\mathbf{C}^s_{i,i'} \;=\; \frac{1}{|\mathcal{O}|}\sum_{j\in\mathcal{O}} \big\|\mathbf{x}_i^{(j)} - \mathbf{x}_{i'}^{(j)}\big\|^2, \qquad \mathbf{C}^t_{k,k'} \;=\; \frac{1}{|\mathcal{O}^c|}\sum_{l\in\mathcal{O}^c} \big\|\mathbf{z}_k^{(l)} - \mathbf{z}_{k'}^{(l)}\big\|^2. \tag{5}$$

This asymmetry makes the GW penalty in eq. equation 1 favor mappings in which similar *observed-side* source pairs correspond to bank pairs with similar *target-side* structure, providing a relational signal beyond feature-only matching. We solve the resulting entropic FGW problem with POT Flamary et al. (2024) per batch to obtain $\widehat{\mathbf{T}}$, then complete every missing modality $l \in \mathcal{O}^c$ via the barycentric projection

$$\hat{\mathbf{x}}_i^{(l)} \;=\; \sum_{k=1}^{K} w^*_{i,k}\,\mathbf{z}_k^{(l)}, \qquad w^*_{i,k} \;:=\; \frac{\widehat{\mathbf{T}}_{i,k}}{\sum_{k'=1}^{K}\widehat{\mathbf{T}}_{i,k'}}. \tag{6}$$

Pseudo-masks for unlabeled records use the same barycentric rule on bank masks at the soft-probability level and are thresholded at 0.5. Per-record uncertainty e_k filters out spatially incoherent completions at bank refresh.

Theoretical Guarantee. We close this section with a finite-sample excess-risk guarantee for one ERM step of an OT-MIMIC training round, treating the memory bank $\mathcal{Z}$ as a fixed anchor set and the transport map $\widehat{T}$ as drawn from a class of bounded Rademacher complexity, fit independently of the records to which it is later applied (precise measurability and sampling conditions are deferred to the detailed algorithm in Appendix B). Let n_c, n_u be the complete and incomplete sample sizes, $L_\phi^{\mathrm{sc}}, L_\phi, B$ the score Lipschitz, composite Lipschitz, and uniform bound of the surrogate ϕ, $\varepsilon_{\mathrm{OT}}$ the FGW modality completion error, $\varepsilon_{\mathrm{rec}}$ the population recoverability error, $\mathrm{Approx}_{\mathcal{F}}$ the approximation bias of the ERM hypothesis class $\mathcal{F}$, and $\mathfrak{R}_n(\cdot)$ a Rademacher complexity over n iid records.

Theorem 1 (informal, with exact constants and proof in Appendix B, Theorem 2)**.** *Suppose ϕ is bounded and score-Lipschitz, and $\widehat{T}$ comes from a function class of bounded Rademacher complexity, fit on records disjoint from those used in ERM (this covers OT-map distillation into a parametric NN, kernel/ridge regression of the barycentric weights, and similar parametric strategies). Then, with probability at least $1-\delta$ over the records used in ERM, the resulting ERM minimizer $\hat{f}$ satisfies*

$$\begin{aligned} R_\phi(\hat{f}) - R_\phi^\star \;\le\; & \underbrace{L_\phi^{\mathrm{sc}}\big(\mathfrak{R}_{n_c}(\widetilde{\mathcal{F}}^c) + \mathfrak{R}_{n_u}(\widetilde{\mathcal{F}}_{\widehat{T}})\big) + B\sqrt{\tfrac{2\log(2/\delta)}{n_c}} + B\sqrt{\tfrac{2\log(2/\delta)}{n_u}}}_{\textit{ERM concentration}} \\ & + \underbrace{L_\phi \varepsilon_{\mathrm{OT}}}_{\textit{OT}} + \underbrace{L_\phi \varepsilon_{\mathrm{rec}}}_{\textit{recoverability}} + \underbrace{\mathrm{Approx}_{\mathcal{F}}}_{\textit{approx.}}, \end{aligned} \tag{7}$$

where $\widetilde{\mathcal{F}}^c, \widetilde{\mathcal{F}}_{\widehat{T}}$ are the induced complete-stratum and completion-composed function classes defined in Appendix B. Theorem 2 in the Appendix states the version with sharp constants.

Bound interpretation and algorithmic design. The four error sources in eq. equation 7 are not bookkeeping. Each one motivates a concrete OT-MIMIC design choice. **(i) ERM concentration** decays at $\tilde{O}(1/\sqrt{n_c} + 1/\sqrt{n_u})$ for standard NN classes, supporting the use of *all* samples after FGW modality completion rather than discarding incomplete records. **(ii) The OT term** $\varepsilon_{\mathrm{OT}}$ is the quantity most directly affected by bank quality, motivating an *uncertainty-curated* memory bank (favoring high-confidence anchors at refresh) over a generic matching cache. **(iii) Recoverability** $\varepsilon_{\mathrm{rec}}$ captures

population-level information loss. Our asymmetric FGW cost in eq. equation 5 is designed to exploit approximate cross-modal isometry between observed-side and complementary-side structure rather than relying on feature similarity alone. **(iv) Approximation** $\mathrm{Approx}_{\mathcal{F}}$ is the usual hypothesis-class bias, controlled by the segmentation backbone (a ConvNeXt-based U-Net with language-guided fusion in our instantiation). Appendix B.5 gives the term-by-term reading and an objective-level condition under which the FGW structural term provides information beyond feature-only Wasserstein matching.

Theorem–practice gap. The deployed POT solver remains outside Theorem 1: its empirical FGW projection is non-parametric, and the same source batch is used for completion and ERM. To partially bridge this gap, we additionally evaluate a bounded-complexity neural entropic FGW estimator that predicts barycentric transport weights from the observed query and a fixed memory bank. Under the stated one-epoch double-sample-split assumptions, this parametric variant falls within the theorem's function class, whereas the empirically stronger multi-epoch POT implementation remains outside the guarantee. Appendix B.6 details the theoretical gap, and full POT-versus-neural results are reported in Appendix G.

4 Experiments

4.1 Experimental Setup

Datasets and Implementation. We evaluate the proposed framework on two publicly available medical segmentation benchmarks: **QaTa-COV19** Yamac et al. (2021) (9,258 chest X-rays with pixel-level lesion masks) and **MosMedData+** Morozov et al. (2020) (2,729 CT slices of pulmonary infections), with textual descriptions and official splits from Li et al. (2023). Images and masks are resized to 224×224 and channel-wise normalized. We train in PyTorch on a single H100 with AdamW (lr 1×10^{-4}, weight decay 1×10^{-2}, batch 32, cosine annealing to 10^{-6}, early stopping at 20-epoch patience, max 50 epochs). Full setup, splits, and reproducibility notes are in Appendix E.

4.2 Comparison Experiments

Setup, data access, and baselines. For each $(\text{dataset}, \text{ratio})$ cell, we synthesize modality imbalance by uniformly at random withholding one modality from a fraction $1 - r$ of training samples while keeping the complementary modality intact, where $r \in \{50\%, 25\%, 10\%, 5\%, 1\%\}$ denotes the fraction of fully paired image, text, and mask training samples. We consider two scenarios: **missing-text**, where text is removed while the image and mask remain available, and **missing-image-and-mask**, where the image and mask are jointly removed while text remains available. Validation and test sets are fully paired for all methods, so missingness occurs only during training. All baselines see the same training set as OT-MIMIC. The difference lies in how they use incomplete records. LViT Li et al. (2023) and GuideSeg Zhong et al. (2023) cannot natively consume samples with missing image or text inputs and therefore discard those incomplete records during training. LViT can, however, use samples with missing masks. ProLearn Ye et al. (2025) handles missing text through prototype matching. OT-MIMIC handles all six patterns in eq. equation 2 through transport-based barycentric completion. The "Handle" columns of Tables 1–2 indicate which types of missingness each baseline can natively handle.

Additional comparisons. Appendix D reports supervised image-only benchmarking. Appendix G further gives fixed-pipeline completion ablations, a theory-compatible neural FGW estimator, MedCLIPSeg and DuSSS comparisons, Kvasir-SEG cross-domain results, a weakened-text diagnostic, and compute profiles.

Main Results. Tables 1–2 report fixed-seed MCAR experiments on two thoracic datasets. The coarse spatial and laterality cues in the LViT-derived text are detailed in Appendix E and stress-tested in Appendix G.5. OT-MIMIC matches or exceeds the strongest baseline in all ten missing-text cells and in eight of ten missing-image-and-mask cells. At only 1% complete-pair availability, its Dice gain over the strongest baseline is 3.1 points on QaTa-COV19 and 5.1 points on MosMedData+ in the missing-text setting. In the missing-image-and-mask setting, the corresponding gains are 6.8 and 5.4 points. These results support the broader strategy of retaining partially observed records

Dataset	Model	Handle			50%		25%		10%		5%		1%	
		T	I	M	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
QaTa-COV19	LViT			√	0.850	0.739	0.840	0.724	0.801	0.668	0.760	0.613	0.667	0.501
	GuideSeg				0.897	0.813	0.888	0.798	0.879	0.784	0.862	0.757	0.732	0.578
	ProLearn	√			0.867	0.772	0.860	0.769	0.858	0.757	0.857	0.756	0.857	0.755
	OT-MIMIC	√	√	√	**0.908**	**0.831**	**0.908**	**0.831**	**0.898**	**0.814**	**0.895**	**0.809**	**0.888**	**0.799**
MosMed Data+	LViT			√	0.727	0.571	0.721	0.564	0.680	0.515	0.573	0.402	0.316	0.188
	GuideSeg				0.748	0.597	0.748	0.598	0.729	0.573	0.679	0.514	0.512	0.344
	ProLearn	√			0.754	0.613	0.751	0.611	0.742	0.609	0.738	0.607	0.722	0.603
	OT-MIMIC	√	√	√	**0.775**	**0.633**	**0.774**	**0.631**	**0.778**	**0.637**	**0.770**	**0.626**	**0.773**	**0.630**

Table 1: **Multimodal segmentation performance under the missing-text setting on QaTa-COV19 and MosMedData+.** We report Dice and mIoU at text availability, or complete-case, ratios of 50%, 25%, 10%, 5%, and 1%. In the **Handle** columns, **T**, **I**, and **M** indicate whether a method can natively handle missing **text**, **image**, and **mask** inputs, respectively.

Dataset	Model	Handle			50%		25%		10%		5%		1%	
		T	I	M	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
QaTa-COV19	LViT			√	0.850	0.739	0.840	0.724	0.801	0.668	0.760	0.613	0.667	0.501
	GuideSeg				0.897	0.813	0.888	0.798	**0.879**	0.784	**0.862**	**0.757**	0.732	0.578
	ProLearn	√			0.863	0.759	0.851	0.740	0.835	0.716	0.799	0.666	0.676	0.511
	OT-MIMIC	√	√	√	**0.907**	**0.830**	**0.903**	**0.823**	**0.879**	**0.785**	0.860	0.755	**0.800**	**0.667**
MosMed Data+	LViT			√	0.727	0.571	0.721	0.564	0.680	0.515	0.573	0.402	0.316	0.188
	GuideSeg				0.748	0.597	0.748	0.598	**0.729**	**0.573**	0.679	0.514	0.512	0.344
	ProLearn	√			0.721	0.564	0.714	0.555	0.658	0.490	0.587	0.415	0.150	0.081
	OT-MIMIC	√	√	√	**0.769**	**0.624**	**0.752**	**0.603**	0.711	0.552	**0.684**	**0.519**	**0.566**	**0.395**

Table 2: **Multimodal segmentation performance under the missing-image-and-mask setting on QaTa-COV19 and MosMedData+.** We report Dice and mIoU at image + mask availability, or complete-case, ratios of 50%, 25%, 10%, 5%, and 1%. **Handle** is defined as in Table 1. Notably, the results of LViT and GuideSeg are identical to those in the missing-text setting, since these baselines cannot handle either missing text or missing image inputs and are therefore evaluated on the same subset of modality-complete data in both settings.

through transport-based completion rather than discarding them. In the missing-text regime, the remaining 99% of records still retain image and mask supervision, so the gain reflects the use of those incomplete records after completion. With the backbone, memory bank, incomplete records, and training schedule fixed, our structure-aware FGW instantiation also outperforms nearest-anchor and feature-mean completion in every QaTa-COV19 cell, as shown in Appendix G.1. At 1% availability, removing lesion-count and location cues reduces Dice by only 1.7 points on QaTa-COV19 and 1.0 point on MosMedData+, while preserving the margins over the baselines. See Appendix G.5. Cross-domain Kvasir-SEG and MedCLIPSeg results are reported in Appendix G.3, while the auxiliary mask-only-missing comparison with DuSSS and LViT is reported in Appendix G.4.

4.3 Ablation Studies

Impact of memory bank capacity. Table 3 evaluates the sensitivity to bank capacity K on the missing-image-and-mask setting, the missing-text counterpart is in Table 5 of the Appendix. Performance is remarkably robust once the bank reaches a moderate size (5%–25% of the training set): even under extreme paired-supervision scarcity (1% availability), a 5% bank is already sufficient, and growing to 25% yields only marginal gains. These results show that OT-MIMIC is stable across moderate bank capacities and does not require exhaustive storage of the training set.

Dataset	Capacity	50%		25%		10%		5%		1%	
		Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
QaTa-COV19	1% (57)	0.906	0.829	0.901	0.820	0.878	0.782	**0.860**	0.754	**0.800**	0.666
	5% (286)	0.906	0.828	**0.903**	**0.823**	**0.879**	0.784	**0.860**	0.754	**0.800**	**0.667**
	10% (572)	**0.908**	**0.831**	**0.903**	**0.823**	**0.879**	**0.785**	**0.860**	**0.755**	**0.800**	**0.667**
	25% (1429)	0.907	0.830	**0.903**	**0.823**	**0.879**	**0.785**	**0.860**	**0.755**	**0.800**	**0.667**
MosMed Data+	1% (22)	0.757	0.609	0.729	0.574	0.680	0.515	0.590	0.419	**0.567**	**0.396**
	5% (109)	0.764	0.618	0.746	0.595	0.706	0.545	**0.684**	0.519	0.566	0.395
	10% (218)	**0.769**	**0.625**	0.746	0.595	**0.711**	**0.552**	**0.684**	**0.520**	0.566	0.395
	25% (546)	**0.769**	0.624	**0.752**	**0.603**	**0.711**	**0.552**	**0.684**	0.519	0.566	0.395

Table 3: **Ablation on memory bank capacity under missing-image-and-mask scenarios on QaTa-COV19 and MosMedData+.** Memory bank capacity is reported as a fraction of the full training set, with the corresponding number of stored samples in parentheses.

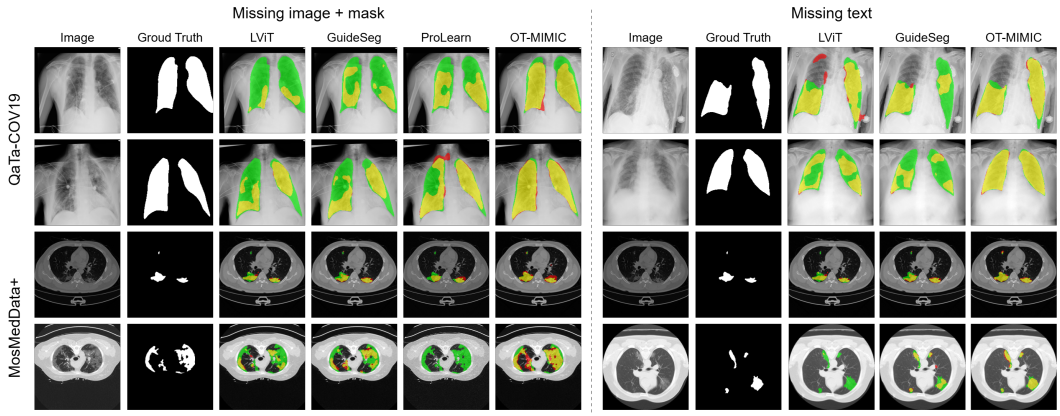

Figure 3: **Qualitative comparisons under different missing modalities.** Predictions under severe missing ratios for both missing-text and missing-image-and-mask settings on QaTa-COV19 and MosMedData+. **Green**, **red**, and **yellow** regions represent the ground truth (GT), model predictions, and their accurate overlap, respectively.

4.4 Qualitative Segmentation Results

Figure 3 visualizes representative predictions under extreme missing ratios. Compared with prior vision–language baselines, OT-MIMIC produces more coherent masks with fewer spurious regions and better boundary adherence on low-contrast lesions and fragmented infection patterns. These examples are consistent with the quantitative improvements observed under severe missingness. Additional cases for both settings are in Appendix J.

5 Conclusion

OT-MIMIC provides a transport-based framework for learning from incomplete image–text records with missing segmentation supervision. It maps partially observed records to a curated memory bank and reuses transport weights for barycentric completion, with asymmetric FGW serving as our structure-aware empirical instantiation. Its excess-risk analysis separates ERM, transport-completion, recoverability, and approximation errors, while the empirical study shows the largest gains under extreme paired-data scarcity. Controlled ablations, recent VLM and semi-supervised baselines, a Kvasir-SEG extension, weakened-text diagnostics, and compute profiles are provided in Appendix G. The theory-compatible neural-estimator variant is evaluated in Appendix G.2. The empirically preferred POT multi-epoch implementation remains outside the theorem. MCAR, statistical, and text-provenance limitations are discussed in Appendix H, and broader impacts in Appendix I.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, improving clarity and readability, drafting and refining portions of the manuscript, literature retrieval and discovery, and the presentation and checking of parts of the theoretical analysis and mathematical proofs. All AI-assisted content was reviewed, verified, and revised by the authors. All experiments reported in this paper were implemented and executed by the authors on the stated datasets and computational setup, and all reported experimental results were obtained from actual model runs and verified by the authors. Generative AI tools were not used to generate synthetic datasets or experimental results. The authors take full responsibility for the final content of the paper, including its methods, theoretical claims, experimental results, analyses, and references.

REPRODUCIBILITY STATEMENT

We provide detailed resources to support reproducibility throughout the paper and appendix. The complete OT-MIMIC training and completion procedures are specified in Algorithms 1 and 2. Appendix B states the theoretical assumptions and provides complete proofs, while Appendix E documents the datasets, preprocessing, model configuration, optimization settings, FGW solver parameters, data splits, and reproducibility details. Additional controlled experiments and diagnostic settings are described in Appendix G. Upon acceptance, we will release the training and evaluation code, exact missingness masks, dataset split identifiers, experiment configurations, checkpoint information, and environment specifications.

REFERENCES

Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pp. rs–3, 2024.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Rémi Flamary, Cédric Vincent-Cuaz, Nicolas Courty, Alexandre Gramfort, Oleksii Kachaiev, Huy Quang Tran, Laurène David, Clément Bonet, Nathan Cassereau, Théo Gnassounou, Eloi Tanguy, Julie Delon, Antoine Collas, Sonia Mazelet, Laetitia Chapel, Tanguy Kerdoncuff, Xizheng Yu, Matthew Feickert, Paul Krzakala, Tianlin Liu, and Eduardo Fernandes Montesuma. Pot python optimal transport (version 0.9.5), 2024. URL `https://github.com/PythonOT/POT`.

Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Hetero-modal image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 469–477. Springer, 2016.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European conference on computer vision*, pp. 108–124. Springer, 2016.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling*, 2020.

Donggeun Kim and Taesup Kim. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In *European Conference on Computer Vision*, pp. 171–187. Springer, 2024.

Taha Koleilat, Hojat Asgariandehkordi, Omid Nejatimanzari, Berardino Barile, Yiming Xiao, and Hassan Rivaz. Medclipseg: Probabilistic vision-language adaptation for data-efficient and generalizable medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1406–1417, 2026.

Aishik Konwer, Xiaoling Hu, Joseph Bae, Xuan Xu, Chao Chen, and Prateek Prasanna. Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21415–21425, 2023.

Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, Yifan Liu, Zhen Chen, and Yixuan Yuan. U-kan makes strong backbone for medical image segmentation and generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4652–4660, 2025.

Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*, 43(1):96–107, 2023.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

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, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 2302–2310, 2021.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

Sergey P Morozov, Anna E Andreychenko, Nikolay A Pavlov, AV Vladzymyrskyy, Natalya V Ledikhova, Victor A Gombolevskiy, Ivan A Blokhin, Pavel B Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Qingtao Pan, Wenhao Qiao, Jingjiao Lou, Bing Ji, and Shuo Li. Dusss: Dual semantic similarity-supervised vision-language model for semi-supervised medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6299–6307, 2025. doi: 10.1609/aaai.v39i6.32674.

Jinzhao Qian, Ankita Joshi, Hailong Li, Nehal A Parikh, Jonanthan R Dillman, and Lili He. Utility of deep learning to address missing modalities from multi-modal medical imaging studies: A systematic review. In *Artificial Intelligence and Applications*, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Junjie Shi, Caozhi Shang, Zhaobin Sun, Li Yu, Xin Yang, and Zengqiang Yan. Passion: Towards effective incomplete multi-modal medical image segmentation with imbalanced missing rates. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 456–465, 2024.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

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. IEEE, 2023a.

Tao Wang and Ziv Goldfeld. Neural entropic optimal transport and gromov–wasserstein alignment. *IEEE Transactions on Information Theory*, 72(4):2424–2443, 2026. doi: 10.1109/TIT.2026.3661439.

Yixin Wang, Yang Zhang, Yang Liu, Zihao Lin, Jiang Tian, Cheng Zhong, Zhongchao Shi, Jianping Fan, and Zhiqiang He. Acn: adversarial co-training network for brain tumor segmentation with missing modalities. In *International conference on medical image computing and computer-assisted intervention*, pp. 410–420. Springer, 2021.

Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22025–22034, 2023b.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, pp. 3876, 2022.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*, 2024.

Mehmet Yamac, Mete Ahishali, Aysen Degerli, Serkan Kiranyaz, Muhammad EH Chowdhury, and Moncef Gabbouj. Convolutional sparse support estimator-based covid-19 recognition from x-ray images. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):1810–1820, 2021.

Qiushi Yang, Xiaoqing Guo, Zhen Chen, Peter YM Woo, and Yixuan Yuan. D 2-net: Dual disentanglement network for brain tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, 41(10):2953–2964, 2022a.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18155–18165, 2022b.

Shuchang Ye, Usman Naseem, Mingyuan Meng, and Jinman Kim. Alleviating textual reliance in medical language-guided segmentation via prototype-driven semantic approximation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22316–22326, 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 107–117. Springer, 2022.

Zechen Zhao, Heran Yang, and Jian Sun. Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In *International conference on medical image computing and computer-assisted intervention*, pp. 183–192. Springer, 2022.

Yi Zhong, Mengqiu Xu, Kongming Liang, Kaixin Chen, and Ming Wu. Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 724–733. Springer, 2023.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pp. 3–11. Springer, 2018.

APPENDIX

Appendix A. Algorithm 14
Appendix B. Theory Proofs 14
Appendix C. Representation-Level Comparison of OT Variants 20
Appendix D. Image-only Comparison 20
Appendix E. Experimental Details 21
Appendix F. Ablation Study 23
Appendix G. Additional Controlled, Cross-Domain, and Diagnostic Experiments 23
Appendix H. Limitations 26
Appendix I. Broader Impacts 27
Appendix J. More Visualization 27

A ALGORITHM

For completeness, we provide a modular pseudocode description of OT-MIMIC. Algorithm 1 describes the overall training pipeline, including segmentation updates, uncertainty-based memory-bank refinement, and iterative completion of incomplete samples. Algorithm 2 details the FGW-based modality completion step used to complete missing modalities and produce pseudo-labels for unlabeled cases. We use $\widetilde{\mathcal{D}}$ to denote the completed training set/representations after each round of modality completion.

Algorithm 1: OT-MIMIC Training

Input: Incomplete training set $\mathcal{D}$; memory bank capacity K; FGW hyperparameters α, ε
Output: Model parameters θ

Initialize memory bank $\mathcal{Z}$ with complete cases
▷ reliable anchors for OT modality completion
Initialize completed dataset/representations $\widetilde{\mathcal{D}} \leftarrow \text{FGWCOMPLETE}(\mathcal{D}, \mathcal{Z}, \alpha, \varepsilon)$
for $t = 1, 2, \ldots, T$ **do**
  Update θ on $\widetilde{\mathcal{D}}$
  Compute per-sample segmentation cross-entropy losses $\{\ell_i^{(t)}\}_{i=1}^n$
  ▷ uncertainty / confidence score
  Extract projected tuples $(z_i^{img,t}, z_i^{txt,t}, \hat{z}_i^{mask,t})$
  Update memory bank $\mathcal{Z}$ with high-confidence tuples according to low-loss ranking
  ▷ priority queue, retain top-K entries
  $\widetilde{\mathcal{D}} \leftarrow \text{FGWCOMPLETE}(\mathcal{D}, \mathcal{Z}, \alpha, \varepsilon)$
return θ

B THEORETICAL ANALYSIS: ASSUMPTIONS AND PROOFS

This appendix gives the formal setup, assumptions, lemmas, and the main theorem (cf. Theorem 1). We treat the OT-error term $\varepsilon_{\mathrm{OT}}$ as an abstract conditional perturbation rather than deriving a specific rate. Appendix B.6 discusses how the implemented POT-based estimator fits the framework.

B.1 SETUP

Data-generation process. We state the proof for two input modalities, the smallest setting that contains the statistical difficulty of OT-MIMIC: one modality observed for everyone, a second modality missing for a fraction of records, and a downstream label. Extending to more modalities or to several missing modalities only changes the bookkeeping, the OT/ERM logic is identical. Let (A_i, B_i, Y_i), $i = 1, \ldots, n$, be drawn from a population distribution P after passing the raw data through a *frozen* pretrained representation extractor (e.g., BiomedCLIP). Here $A_i \in \mathcal{A}$ is modality 1 (observed for all i), $B_i \in \mathcal{B}$ is modality 2, and $Y_i \in \mathcal{Y}$ is the segmentation mask.

Algorithm 2: FGWCOMPLETE$(\mathcal{D}, \mathcal{Z}, \alpha, \varepsilon)$

Input: Incomplete training set $\mathcal{D}$; memory bank $\mathcal{Z}$; FGW hyperparameters α, ε

Output: Completed dataset/representations $\widetilde{\mathcal{D}}$

For each sample i, identify the observed modality set $\mathcal{O}_i \subseteq \{1, \dots, M\}$
▷ M: total number of modalities

Extract projected observed-modality representations $\mathbf{x}_i^{(r)} = g_r(f_r(x_i^{(r)}))$, for $r \in \mathcal{O}_i$
▷ f_r: encoder; g_r: projection module

Read memory embeddings $\{\mathbf{z}_k^{(r)}\}_{k=1,r=1}^{K,M}$ from $\mathcal{Z}$
▷ k: memory-record index; r: modality index

Build the cross-domain ground cost matrix $\mathbf{M}$ and the source/target pairwise distance matrices $(\mathbf{C}_s, \mathbf{C}_t)$
▷ s: query/source side; t: memory-bank/target side

Solve the entropic FGW problem defined in Eq. 1
▷ obtain the empirical transport plan $\widehat{\mathbf{T}}$

Complete each missing modality $l' \notin \mathcal{O}_i$ using transport-weighted memory-bank aggregation

Generate pseudo-labels $\hat{y}_i$ for unlabeled samples when needed

return $\widetilde{\mathcal{D}}$

The observation pattern is a *fixed-count stratified design*: with $p := \mathbb{P}(O{=}1) \in (0,1)$ the population complete-case probability, fix complete budget n_c and missing budget $n_u = n - n_c$ and a fixed split $I_c \sqcup I_u = \{1, \dots, n\}$. Independently draw n_c iid tuples from $P(\cdot \mid O{=}1)$ (assigned to I_c) and n_u iid tuples from $P(\cdot \mid O{=}0)$ (assigned to I_u, with B_i unobserved). The two strata are mutually independent by construction.

One-epoch algorithm with double sample-splitting. The theoretical algorithm is a single-pass, train-and-freeze idealization of the adaptive memory-bank scheme. Two ingredients are committed up front: (i) the completion rule is *NN-parametrized barycentric over the bank* (Assumption D); (ii) *both* strata are split into halves, with one half feeding the completion-rule fitter and the other feeding ERM. Concretely, fix even splits $I_c = I_c^{(1)} \sqcup I_c^{(2)}$ with $|I_c^{(j)}| = m_c := n_c/2$ and $I_u = I_u^{(1)} \sqcup I_u^{(2)}$ with $|I_u^{(j)}| = m := n_u/2$. Let $S_u^{(j)} := ((A_i, Y_i, H_i))_{i \in I_u^{(j)}}$, $\widetilde{S}_u^{(j)} := ((A_i, B_i, Y_i, H_i))_{i \in I_u^{(j)}}$, $\mathcal{M} := ((A_i, B_i, Y_i, H_i))_{i \in I_c^{(1)}}$ (the OT anchor), and $\widetilde{\mathcal{M}}^{(2)} := ((A_i, B_i, Y_i))_{i \in I_c^{(2)}}$ (the ERM-side complete records). Set $n_e := m_c + m$.

(1) Freeze the encoder.
(2) Form the OT-anchor memory bank $\mathcal{M}$ from the first-half complete records ($I_c^{(1)}$, size m_c).
(3) Fit the NN-parametrized barycentric weights $w_{\widehat{\theta}}$ on $(\mathcal{M}, S_u^{(1)})$ by parametric FGW minimization (or distillation from an empirical FGW coupling on $S_u^{(1)} \times \mathcal{M}$); set $\widehat{T}(h) := \sum_{k=1}^{m_c} B_k \, w_{\widehat{\theta}}(h)_k$.
(4) Complete the missing B_i, $i \in I_u^{(2)}$, by $\widehat{T}(H_i)$.
(5) Run ERM *once* on the held-out halves $I_c^{(2)} \cup I_u^{(2)}$ (size n_e): second-half complete records keep modality 2; second-half incomplete records use the completed one.
(6) The bank is *not* updated during the ERM pass; the first halves $I_c^{(1)}$ and $I_u^{(1)}$ are consumed in fitting $\widehat{T}$ and not reused in ERM.

Double sample-splitting makes both stratum-internal Rademacher complexities applicable as fixed-class quantities: conditional on $(\mathcal{M}, S_u^{(1)})$, $\widehat{T}$ is fixed and the ERM data is iid and independent of $\widehat{T}$.

Notation. Set $W_i := (A_i, H_i, O_i)$ where the *observed block* H_i equals A_i in the input-only analysis (or (A_i, Y_i) if the mask is treated as another observed modality at training time). Let $T^\star : \mathcal{H} \to \mathcal{B}$ be

a fixed (non-random) population reference completion map. Define

$$\Psi^\star(W,B) := \begin{cases}(A,B,O), & O=1,\\ (A,T^\star(H),O), & O=0,\end{cases} \quad \widehat{\Psi}(W,B) := \begin{cases}(A,B,O), & O=1,\\ (A,\widehat{T}(H),O), & O=0,\end{cases}$$

and $\Psi^{\mathrm{full}}(A,B,O) := (A,B,O)$. All completed representations live in a metric space $(\mathcal{Z}, d_{\mathcal{Z}})$. Let $\mathcal{F}$ be a class of measurable real-valued predictors on $\mathcal{Z}$ and $\phi : \mathcal{Y}\times\mathbb{R}\to[0,\infty)$ a surrogate loss. Because $\widehat{T}$ is random we distinguish the conditional population risk $R_\phi^{\mathcal{M},S_u}(f;\widehat{T}) := \mathbb{E}[\phi(Y, f(\widehat{\Psi}(W,B))) \mid \mathcal{M}, S_u]$ from the deterministic reference $R_\phi(f;T^\star) := \mathbb{E}[\phi(Y, f(\Psi^\star(W,B)))]$ and the no-missing benchmark $R_\phi^{\mathrm{full}}(f) := \mathbb{E}[\phi(Y, f(\Psi^{\mathrm{full}}(A,B,O)))]$, with Bayes risk $R_\phi^\star := \inf_{f \text{ meas.}} R_\phi^{\mathrm{full}}(f)$. The expectation in $R_\phi^{\mathcal{M},S_u}$ is over a fresh draw $(A,B,Y,O)\sim P$, independent of $(\mathcal{M}, S_u)$.

B.2 Assumptions

Assumption A (Bounded surrogate)**.** *There exists $B<\infty$ with $0\le\phi(y,f(z))\le B$ for every $y\in\mathcal{Y}$, $f\in\mathcal{F}$, and $z\in\{\widehat{\Psi}(W,B), \Psi^\star(W,B), \Psi^{\mathrm{full}}(A,B,O)\}$.*

Assumption B (Composite and score Lipschitzness)**.** (i) Composite: $|\phi(y,f(z))-\phi(y,f(z'))|\le L_\phi\, d_{\mathcal{Z}}(z,z')$ *for all $f\in\mathcal{F}$, y, z, z'.* (ii) Score: $|\phi(y,t)-\phi(y,t')|\le L_\phi^{\mathrm{sc}}|t-t'|$ *for all y, t, t'. The two are linked by $L_\phi = L_\phi^{\mathrm{sc}}\cdot L_f$ when each $f\in\mathcal{F}$ is L_f-Lipschitz w.r.t. $d_{\mathcal{Z}}$.*

Assumption C (Measurability and non-degeneracy)**.** (i) $p=\mathbb{P}(O{=}1)\in(0,1)$ *and* $n_c, n_u\ge 1$. (ii) $T^\star$ *and* ϕ *are Borel; there is a Borel* assembly map $\rho : \mathcal{A}\times\mathcal{B}\times\{0,1\}\to\mathcal{Z}$ *such that* $\Psi^{\mathrm{full}}(A,B,O)=\rho(A,B,O)$ *etc.* (iii) *The completion rule $(\mathcal{M}, S_u^{(1)}, h)\mapsto\widehat{T}(h)$ and the induced evaluation map are Borel measurable.* (iv) *$\mathcal{F}$ is pointwise measurable or separable, and $\widehat{f}$ is selected by a measurable rule.*

Assumption D (Parametric barycentric modality completion rule)**.** *The completion rule has the barycentric form $\widehat{T}(h)=\sum_{k=1}^{m_c} B_k\, w(h)_k$, where the soft-assignment vector $w(h)\in\Delta^{m_c-1}$ is produced by a parametric map $w_\theta : \mathcal{H}\to\Delta^{m_c-1}$ fitted on $(\mathcal{M}, S_u^{(1)})$ alone (e.g., a small NN distilled from an FGW solver's barycentric weights, or kernel/ridge regression on the same), and the chosen function class $\{w_\theta : \theta\in\Theta\}$ has bounded Rademacher complexity. The realized $\widehat{T}$ takes values in* $\mathrm{conv}(\{B_k : k\in I_c^{(1)}\})$. *The empirical entropic FGW barycentric projection produced by a generic non-parametric solver such as POT is* not *covered by this assumption as a stand-alone completion rule, because it lacks an a priori function-class bound; we treat that case separately in Appendix B.6.*

B.3 Definitions and Lemmas

Definition 1 (ERM-half empirical risks and Rademacher complexities)**.** *Define*

$$\widehat{R}_c(f) := \tfrac{1}{m_c}\textstyle\sum_{i\in I_c^{(2)}}\phi(Y_i, f(A_i,B_i,1)), \qquad \bar{R}_c(f) := \mathbb{E}[\phi(Y, f(A,B,1)) \mid O{=}1],$$
$$\widehat{R}_u(f;\widehat{T}) := \tfrac{1}{m}\textstyle\sum_{i\in I_u^{(2)}}\phi(Y_i, f(A_i,\widehat{T}(H_i),0)), \quad \bar{R}_u(f;\widehat{T}) := \mathbb{E}[\phi(Y, f(A,\widehat{T}(H),0)) \mid O{=}0, \mathcal{M}, S_u^{(1)}],$$

together with loss classes $\mathcal{F}^c := \{(A,B,Y)\mapsto\phi(Y,f(A,B,1)) : f\in\mathcal{F}\}$ (fixed) and $\mathcal{G}_{\widehat{T}} := \{(A,H,Y)\mapsto\phi(Y,f(A,\widehat{T}(H),0)) : f\in\mathcal{F}\}$ (fixed conditional on $(\mathcal{M}, S_u^{(1)})$), and their population Rademacher complexities $\mathfrak{R}_{m_c}(\mathcal{F}^c)$, $\mathfrak{R}_m(\mathcal{G}_{\widehat{T}})$ over $\sigma_i \overset{\mathrm{iid}}{\sim}\mathrm{Unif}\{\pm1\}$ and iid records.

Definition 2 (Recoverability and OT estimation errors)**.** $\varepsilon_{\mathrm{rec}} := \mathbb{E}[d_{\mathcal{Z}}\big(\Psi^\star(W,B), \Psi^{\mathrm{full}}(A,B,O)\big)] = \mathbb{P}(O{=}0)\,\mathbb{E}[d_{\mathcal{Z}}((A,T^\star(H),0),(A,B,0)) \mid O{=}0]$ *and* $\varepsilon_{\mathrm{OT}} := \mathbb{E}[d_{\mathcal{Z}}\Big(\widehat{\Psi}(W,B), \Psi^\star(W,B)\Big) \mid \mathcal{M}, S_u] = \mathbb{P}(O{=}0)\,\mathbb{E}[d_{\mathcal{Z}}\Big((A,\widehat{T}(H),0),(A,T^\star(H),0)\Big) \mid O{=}0, \mathcal{M}, S_u]$.

Lemma 1 (Symmetrization on the ERM halves)**.** *Fix the bank $\mathcal{M}$ and the auxiliary records $S_u^{(1)}$ used to fit $\widehat{T}$. Conditional on $(\mathcal{M}, S_u^{(1)})$, the loss classes $\mathcal{F}^c$ and $\mathcal{G}_{\widehat{T}}$ of Definition 1 are deterministic*

and uniformly bounded by B (Assumption A), and the ERM-half samples $\{(A_i, B_i, Y_i)\}_{i\in I_c^{(2)}}$ and $\{(A_i, H_i, Y_i)\}_{i\in I_u^{(2)}}$ are iid (Assumption C). Then

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\widehat{R}_c(f) - \bar{R}_c(f)\right| \,\middle|\, \mathcal{M}, S_u^{(1)}\right] \;\le\; 2\,\mathfrak{R}_{m_c}(\mathcal{F}^c), \tag{8}$$

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\widehat{R}_u(f;\widehat{T}) - \bar{R}_u(f;\widehat{T})\right| \,\middle|\, \mathcal{M}, S_u^{(1)}\right] \;\le\; 2\,\mathfrak{R}_m(\mathcal{G}_{\widehat{T}}), \tag{9}$$

where the outer expectation is over the corresponding ERM-half sample.

Proof. We prove equation 8; the proof of equation 9 is identical with $\mathcal{G}_{\widehat{T}}$ in place of $\mathcal{F}^c$ (the conditioning on $(\mathcal{M}, S_u^{(1)})$ freezes $\widehat{T}$, so $\mathcal{G}_{\widehat{T}}$ is also deterministic). Write $g_f(A, B, Y) := \phi(Y, f(A, B, 1)) \in \mathcal{F}^c$ and let $S = \{(A_i, B_i, Y_i)\}_{i\in I_c^{(2)}}$. Introduce an iid ghost copy $S' = \{(A_i', B_i', Y_i')\}_{i\in I_c^{(2)}}$ and Rademacher signs $(\sigma_i) \overset{\text{iid}}{\sim} \mathrm{Unif}\{\pm 1\}$, both independent of S and of each other. Since $\bar{R}_c(f) = \mathbb{E}_{S'}[\frac{1}{m_c}\sum_i g_f(S_i')]$, Jensen's inequality gives

$$\mathbb{E}_S\Big[\sup_f\left|\widehat{R}_c(f) - \bar{R}_c(f)\right|\Big] \;\le\; \mathbb{E}_{S,S'}\Big[\sup_f\left|\tfrac{1}{m_c}\sum_{i\in I_c^{(2)}}\big(g_f(S_i) - g_f(S_i')\big)\right|\Big].$$

Because $g_f(S_i) - g_f(S_i')$ is symmetric under $S_i \leftrightarrow S_i'$, multiplying each summand by σ_i preserves the joint distribution; the triangle inequality and the definition of Rademacher complexity then yield

$$\mathbb{E}_{S,S',\sigma}\Big[\sup_f\left|\tfrac{1}{m_c}\sum_i \sigma_i\big(g_f(S_i) - g_f(S_i')\big)\right|\Big] \;\le\; 2\,\mathbb{E}_{S,\sigma}\Big[\sup_f\left|\tfrac{1}{m_c}\sum_i \sigma_i g_f(S_i)\right|\Big] \;=\; 2\,\mathfrak{R}_{m_c}(\mathcal{F}^c),$$

all under the conditioning on $(\mathcal{M}, S_u^{(1)})$. Combining these two displays gives equation 8. □

Lemma 2 (OT perturbation)**.** *Under Assumption B, conditional on $(\mathcal{M}, S_u)$, $\sup_{f\in\mathcal{F}}\left|R_\phi^{\mathcal{M},S_u}(f;\widehat{T}) - R_\phi(f;T^\star)\right| \le L_\phi\varepsilon_{\mathrm{OT}}$, hence $\inf_f R_\phi^{\mathcal{M},S_u}(f;\widehat{T}) \le \inf_f R_\phi(f;T^\star) + L_\phi\varepsilon_{\mathrm{OT}}$.*

Proof. For every f, by Jensen's inequality and Assumption B, $\left|R_\phi^{\mathcal{M},S_u}(f;\widehat{T}) - R_\phi(f;T^\star)\right| \le L_\phi\mathbb{E}[d_{\mathcal{Z}}\Big(\widehat{\Psi}(W,B), \Psi^\star(W,B)\Big) \mid \mathcal{M}, S_u] = L_\phi\varepsilon_{\mathrm{OT}}$. □

Lemma 3 (Recoverability perturbation)**.** *Under Assumption B, $\sup_{f\in\mathcal{F}}\left|R_\phi(f;T^\star) - R_\phi^{\mathrm{full}}(f)\right| \le L_\phi\varepsilon_{\mathrm{rec}}$.*

Proof. Identical to Lemma 2, with the joint expectation over $(W, B, Y, O) \sim P$ and $d_{\mathcal{Z}}\big(\Psi^\star, \Psi^{\mathrm{full}}\big)$ (no conditioning needed because $T^\star$ is non-random). □

B.4 Main Theorem

Theorem 2 (Conditional one-epoch excess surrogate-risk bound, double sample-split)**.** *Under Assumptions A to D, let $\widehat{T}$ be the parametric barycentric modality completion rule of Assumption D, fit on $(\mathcal{M}, S_u^{(1)})$ and drawn from a function class of bounded Rademacher complexity. Set $m_c := n_c/2$, $m := n_u/2$, $n_e := m_c + m$. Suppose an ERM minimizer $\widehat{f} \in \arg\min_{f\in\mathcal{F}}\frac{1}{n_e}\sum_{i\in I_c^{(2)}\cup I_u^{(2)}}\phi(Y_i, f(\widehat{\Psi}(W_i, B_i)))$ exists, and let $\mathrm{Approx}_{\mathcal{F}} := \inf_{f\in\mathcal{F}} R_\phi^{\mathrm{full}}(f) - R_\phi^\star$, $\Delta_w := |p - m_c/n_e|$. Conditionally on $(\mathcal{M}, S_u^{(1)})$, with probability at least $1-\delta$ over $(\widetilde{\mathcal{M}}^{(2)}, \widetilde{S}_u^{(2)})$,*

$$\begin{aligned} R_\phi^{\mathcal{M},S_u^{(1)}}(\widehat{f};\widehat{T}) - R_\phi^\star \;\le\; & \tfrac{2n_c}{n_e}L_\phi^{\mathrm{sc}}\mathfrak{R}_{m_c}(\widetilde{\mathcal{F}}^c) + \tfrac{n_cB}{2n_e}\sqrt{\tfrac{2\log(2/\delta)}{m_c}} \\ & + 4L_\phi^{\mathrm{sc}}\mathfrak{R}_m(\widetilde{\mathcal{F}}_{\widehat{T}}) + B\sqrt{\tfrac{2\log(2/\delta)}{m}} \\ & + 2B\Delta_w + L_\phi\varepsilon_{\mathrm{OT}} + L_\phi\varepsilon_{\mathrm{rec}} + \mathrm{Approx}_{\mathcal{F}}, \end{aligned} \tag{10}$$

where $\widetilde{\mathcal{F}}^c := \{f(A, B, 1) : f \in \mathcal{F}\}$, $\widetilde{\mathcal{F}}_{\widehat{T}} := \{f(A, \widehat{T}(H), 0) : f \in \mathcal{F}\}$.

Proof. All conditioning is on $(\mathcal{M}, S_u^{(1)})$; this freezes $\widehat{T}$. The randomness is over $(\widetilde{\mathcal{M}}^{(2)}, \widetilde{S}_u^{(2)})$.

Step 0 (Decomposition).

$$R_\phi^{\mathcal{M},S_u^{(1)}}(\widehat{f};\widehat{T}) - R_\phi^\star = \underbrace{[R_\phi^{\mathcal{M},S_u^{(1)}}(\widehat{f};\widehat{T}) - \inf_f R_\phi^{\mathcal{M},S_u^{(1)}}(f;\widehat{T})]}_{\text{(I) ERM gap}} + \underbrace{[\inf_f R_\phi^{\mathcal{M},S_u^{(1)}}(f;\widehat{T}) - \inf_f R_\phi(f;T^\star)]}_{\text{(II) OT plug-in}}$$
$$+ \underbrace{[\inf_f R_\phi(f;T^\star) - \inf_f R_\phi^{\text{full}}(f)]}_{\text{(III) Recoverability}} + \underbrace{[\inf_f R_\phi^{\text{full}}(f) - R_\phi^\star]}_{\text{(IV) Approximation}}. \tag{11}$$

Pieces (II)–(IV) follow directly from Lemmas 2 and 3 and the definition of $\mathrm{Approx}_{\mathcal{F}}$: $(II) \le L_\phi \varepsilon_{\mathrm{OT}}$, $(III) \le L_\phi \varepsilon_{\mathrm{rec}}$, $(IV) = \mathrm{Approx}_{\mathcal{F}}$. We focus on Piece (I).

Step 1 (Symmetrization). Pick $f^\star \in \arg\min R_\phi^{\mathcal{M},S_u^{(1)}}(\cdot;\widehat{T})$. ERM optimality gives $\widehat{R}_\phi(\widehat{f};\widehat{T}) \le \widehat{R}_\phi(f^\star;\widehat{T})$, so (I) $\le 2\sup_{f\in\mathcal{F}} \left| R_\phi^{\mathcal{M},S_u^{(1)}}(f;\widehat{T}) - \widehat{R}_\phi(f;\widehat{T}) \right|$.

Step 2 (Stratum decomposition). Define $R_\phi^{(n_e)}(f;\widehat{T}) := \frac{m_c}{n_e}\bar{R}_c(f) + \frac{m}{n_e}\bar{R}_u(f;\widehat{T})$. The triangle inequality gives $\left| R_\phi^{\mathcal{M},S_u^{(1)}}(f;\widehat{T}) - \widehat{R}_\phi(f;\widehat{T}) \right| \le$ (W) + (E), with **(W) Weight mismatch:** $R_\phi^{\mathcal{M},S_u^{(1)}}(f;\widehat{T}) - R_\phi^{(n_e)}(f;\widehat{T}) = (p - m_c/n_e)(\bar{R}_c(f) - \bar{R}_u(f;\widehat{T}))$, so $\sup_f$ (W) $\le B\Delta_w$. **(E) Empirical-vs-population:** $\sup_f$ (E) $\le \frac{m_c}{n_e}\Xi_c(\widetilde{\mathcal{M}}^{(2)}) + \Xi_u(\widetilde{S}_u^{(2)})$, where Ξ_c, Ξ_u are the suprema of the stratum-wise empirical-process gaps.

Step 3 (Concentration via McDiarmid + Talagrand). Conditional on $(\mathcal{M}, S_u^{(1)})$, $\widehat{T}$ is fixed. Replacing one record changes a summand of $\widehat{R}_c$ (resp. $\widehat{R}_u$) by at most B/m_c (resp. B/m). Allocating $\delta/2$ to each stratum, McDiarmid gives, with probability $\ge 1 - \delta/2$ each,

$$\Xi_c \le 2\mathfrak{R}_{m_c}(\mathcal{F}^c) + B\sqrt{\tfrac{\log(2/\delta)}{2m_c}}, \quad \Xi_u \le 2\mathfrak{R}_m(\mathcal{G}_{\widehat{T}}) + B\sqrt{\tfrac{\log(2/\delta)}{2m}},$$

using Lemma 1. Talagrand's contraction (with score-Lipschitz Assumption B (ii)) gives $\mathfrak{R}_{m_c}(\mathcal{F}^c) \le L_\phi^{\mathrm{sc}}\mathfrak{R}_{m_c}(\widetilde{\mathcal{F}}^c)$ and $\mathfrak{R}_m(\mathcal{G}_{\widehat{T}}) \le L_\phi^{\mathrm{sc}}\mathfrak{R}_m(\widetilde{\mathcal{F}}_{\widehat{T}})$.

Step 4 (Combine). Plug Steps 1–3 into equation 11: with probability $\ge 1 - \delta$,

$$\text{(I)} \le 2B\Delta_w + \tfrac{4m_c}{n_e}L_\phi^{\mathrm{sc}}\mathfrak{R}_{m_c}(\widetilde{\mathcal{F}}^c) + \tfrac{m_c B}{n_e}\sqrt{\tfrac{2\log(2/\delta)}{m_c}} + 4L_\phi^{\mathrm{sc}}\mathfrak{R}_m(\widetilde{\mathcal{F}}_{\widehat{T}}) + B\sqrt{\tfrac{2\log(2/\delta)}{m}}.$$

Substituting $m_c = n_c/2$ gives $4m_c/n_e = 2n_c/n_e$ and $m_c B/n_e = n_c B/(2n_e)$, yielding equation 10. □

B.5 Term-by-term reading and FGW-vs-Wasserstein sufficient condition

Term-by-term reading of eq. equation 10 / eq. equation 7. The five sub-error sources in Theorem 2 (and its informal restatement, Theorem 1) refine the four-term grouping in the main body and each correspond to a concrete OT-MIMIC design choice. (i) ERM concentration decays at $\tilde{O}(1/\sqrt{n_e})$ for standard NN classes, justifying the use of *all* samples after FGW modality completion rather than discarding incomplete ones; (ii) the stratum-weight mismatch term $2B\Delta_w$ in equation 10 motivates stratified sampling so that the mini-batch composition matches the population complete/incomplete ratio (this is collapsed into the ERM-concentration block of eq. equation 7 via $\Delta_w \to 0$); (iii) the OT error $\varepsilon_{\mathrm{OT}}$ is the only quantity that shrinks with bank quality, motivating an uncertainty-curated memory bank rather than a generic matching cache; (iv) $\varepsilon_{\mathrm{rec}}$ is a population-level information loss, but FGW's GW gauge in eq. equation 1—which our cost design equation 5 aligns with cross-modal isometry between observed-side and complementary-side structure—is the population condition that keeps it small; (v) approximation is the usual hypothesis-class bias, controlled by the choice of segmentation backbone. Appendix B.6 discusses how the multi-epoch adaptive system relaxes the one-epoch idealization while preserving these qualitative connections.

Objective-level condition favoring structure-aware matching. The excess-risk bound itself does not order FGW and feature-only Wasserstein completion; their difference enters through the induced completion and recoverability errors. Let $\rho^s_{i,i'} := \sqrt{\mathbf{C}^s_{i,i'}}$ and $\rho^t_{k,k'} := \sqrt{\mathbf{C}^t_{k,k'}}$ be the observed-side and complementary-side relational distances of eq. equation 5. Suppose there exists a population coupling $\pi^\star$ between the source and bank distributions such that *(a)* $\mathbb{E}_{\pi^\star}[\mathbf{M}_{i,k}] \le \delta_{\mathrm{feat}}$ and *(b)* $\mathbb{E}_{\pi^\star \otimes \pi^\star}[(\rho^s_{i,i'} - \rho^t_{k,k'})^2] \le \delta^2_{\mathrm{iso}}$. Then the population FGW objective evaluated at $\pi^\star$ is at most $(1-\alpha)\delta_{\mathrm{feat}} + \alpha\delta^2_{\mathrm{iso}}$. Thus, when approximate cross-modal isometry holds, FGW has access to a relational signal unavailable to feature-only matching and can favor couplings that are jointly feature- and structure-compatible. This is an objective-level statement only: without additional regularity or identifiability assumptions, it does not imply a quantitative ordering of barycentric completion error or of $\varepsilon_{\mathrm{OT}}$. The Kendall's τ analysis in Figure 4 (Appendix C) provides the corresponding empirical diagnostic: FGW preserves relational structure better than Wasserstein while maintaining comparable feature similarity.

B.6 Theorem–Practice Gap

What the theorem assumes vs. what OT-MIMIC actually does. Theorem 2 treats $\widehat{T}$ as the parametric barycentric modality completion rule of Assumption D: a function $w_\theta : \mathcal{H} \to \Delta^{m_c - 1}$ drawn from a class with bounded Rademacher complexity, fit on $(\mathcal{M}, S^{(1)}_u)$, and applied to held-out source points $S^{(2)}_u$. The implemented OT-MIMIC does *not* fit a parametric model: it solves the empirical entropic FGW problem of eq. equation 1 directly with the POT solver on each batch, against the current bank, and then applies the resulting barycentric projection back to the same batch. This produces two distinct gaps to the theorem.

Gap G1: function-class boundedness. The theorem requires $\widehat{T}$ to lie in a class with bounded Rademacher complexity. The empirical POT barycentric projection does not: each call recomputes a non-convex, solver-dependent stationary point. We therefore do *not* claim that equation 10 applies to POT, and Assumption D explicitly excludes it. To test a theory-compatible realization, we additionally implement a neural entropic transport-plan estimator inspired by Wang & Goldfeld (2026): given the observed query and a frozen bank, a parametric network outputs normalized transport weights and is trained with our entropic FGW objective. Its barycentric completion belongs to the bounded-complexity parametric class assumed by the theorem, so the one-epoch double-sample-split guarantee applies to this variant under the stated assumptions. Appendix G reports its results alongside POT; POT remains empirically stronger in most settings and is therefore our primary implementation.

A concrete reading of $\varepsilon_{\mathrm{OT}}$. For the neural estimator, a useful schematic decomposition is $\varepsilon_{\mathrm{OT}} \lesssim \varepsilon_{\mathrm{plan}} + \varepsilon_{\mathrm{opt}} + \varepsilon_{\mathrm{reg}}$, where the terms denote finite-sample/parametric plan estimation, finite optimization, and entropic-regularization discrepancy, respectively. Neural entropic OT/GW theory provides tools for the first term (Wang & Goldfeld, 2026); for bounded embeddings, plan error transfers to barycentric-completion error. We do not claim a complete end-to-end rate for the fused objective, evolving bank, and multi-epoch training.

Gap G2: independence between completion and ERM data. The theorem requires $\widehat{T}$ to be fit on records disjoint from those used in the ERM step, as enforced by the double sample split $I_c = I^{(1)}_c \sqcup I^{(2)}_c$ and $I_u = I^{(1)}_u \sqcup I^{(2)}_u$ in Section B. The deployed implementation does not enforce this split. At epoch t, the memory bank $\mathcal{M}^{(t-1)}$ used as the FGW target is fixed before the epoch-t optimization and is inherited from the end of epoch $t-1$. This temporal separation prevents the target bank from adapting to the current batch during the FGW solve, but it does not establish the sample independence required by the theorem, since the bank and the epoch-t ERM step can involve the same underlying training records. Moreover, the same epoch-t source points are used as POT inputs and subsequently in the ERM loss. Thus, the deployed POT implementation does not satisfy the double-sample-split assumption. A formally matched implementation would require constructing the completion map from records disjoint from those used for ERM, which we do not impose in the reported experiments.

Why we use POT despite both gaps. In a strictly one-epoch theoretical analysis a parametric NN T_θ is convenient: it gives a smooth, globally defined $\widehat{T}(h)$ with controlled Rademacher complexity. In a multi-epoch implementation, however, the bank changes every epoch, so a NN distilled from one bank would have to be retrained whenever the bank is refreshed. Solving FGW directly with POT against the current bank avoids this re-fitting: the transport plan is recomputed each batch (Sinkhorn cost $O(S_{\mathrm{Sink}}\, bK)$ for a b-point batch and a K-point bank), and naturally tracks the bank. We use POT's entropic FGW solver with $\alpha = 0.5$, $\epsilon = 0.1$, treat the resulting $\widehat{\pi}$ as a non-parametric estimator of the soft-assignment row $w(h) = \widehat{\pi}_{i,\cdot}/\sum_k \widehat{\pi}_{i,k}$, and form $\widehat{T}(h_i) = \sum_k w_{i,k}\mathbf{z}_k$ as the data-driven counterpart of the parametric T_θ analyzed in the proof. Theorem 2 should therefore be read as a guarantee for an idealized (distilled, sample-split) version of OT-MIMIC; the deployed POT-based completion is a practical relaxation whose qualitative error sources are still the four terms of equation 10, but for which a formal bound would require closing G1 (distillation into a parametric class) and the source-point version of G2 (per-epoch sample splitting).

Scope of the theoretical guarantee. The theorem analyzes a single ERM step with a fixed completion map and leaves $\varepsilon_{\mathrm{OT}}$ as an abstract conditional perturbation. It therefore does not model the compounding effects of an evolving multi-epoch bank or provide a numerical rate for the POT completion error. The bound is intended to isolate the statistical contributions of ERM, completion, recoverability, and approximation rather than predict a numerical performance level.

Recoverability and structural matching. $\varepsilon_{\mathrm{rec}}$ measures the irreducible information loss from modality completion with any deterministic reference map $T^\star$ and is non-zero whenever B is not deterministically recoverable from H. The GW summand of FGW in eq. equation 1 favors transport plans that preserve relational geometry between observed-side and complementary-side structures in the current projected space. This provides a structural inductive bias for completion, but it does not by itself guarantee small absolute $\varepsilon_{\mathrm{rec}}$, which ultimately depends on the population data law.

C Representation-Level Comparison of OT Variants

Figure 4 provides a representation-level comparison under the missing-image-and-mask setting with a 50% availability ratio, where observed text is used as the source modality and the memory bank is constructed from complete training cases. Using refined BiomedCLIP embeddings, we compare different OT variants by measuring both cosine similarity to the ground-truth visual representation and Kendall's τ, which reflects whether pairwise structural relations are preserved after transport. Classical Wasserstein transport achieves strong feature-level proximity, with a cosine similarity of 0.5196, but has a lower structural correlation with Kendall's $\tau = 0.5418$. In contrast, Gromov-Wasserstein better preserves relational structure, reaching Kendall's $\tau = 0.5541$, but substantially sacrifices feature-level fidelity, with cosine similarity dropping to 0.2536. FGW achieves the best overall balance, obtaining the highest cosine similarity of 0.5234 and the highest Kendall's τ of 0.5580. Compared with Wasserstein, FGW improves Kendall's τ by 0.0162 while slightly increasing cosine similarity by 0.0038; compared with Gromov-Wasserstein, FGW maintains comparable structural preservation while substantially improving cosine similarity by 0.2698. These results show that FGW provides the strongest balance of feature fidelity and structural preservation among the three transport variants in this representation-level diagnostic.

D Image-only Comparison

To further contextualize our results in Section 4, we compare OT-MIMIC against robust pure-vision segmentation backbones trained on the *full* set of available images without any missing-modality regime. These baselines fall into three families: (i) CNN-based architectures including U-Net Ronneberger et al. (2015), U-Net++ Zhou et al. (2018), nnU-Net Isensee et al. (2021), and Attention U-Net Oktay et al. (2018); (ii) Transformer-based models such as TransU-Net Chen et al. (2021) and Swin-U-Net Cao et al. (2022); and (iii) emerging or foundational methods like U-KAN Li et al. (2025) and MedSAM Ma et al. (2024). This comparison is contextual rather than causal: it situates OT-MIMIC against strong image-only models, but does not isolate the effect of OT because OT-MIMIC also exploits text.

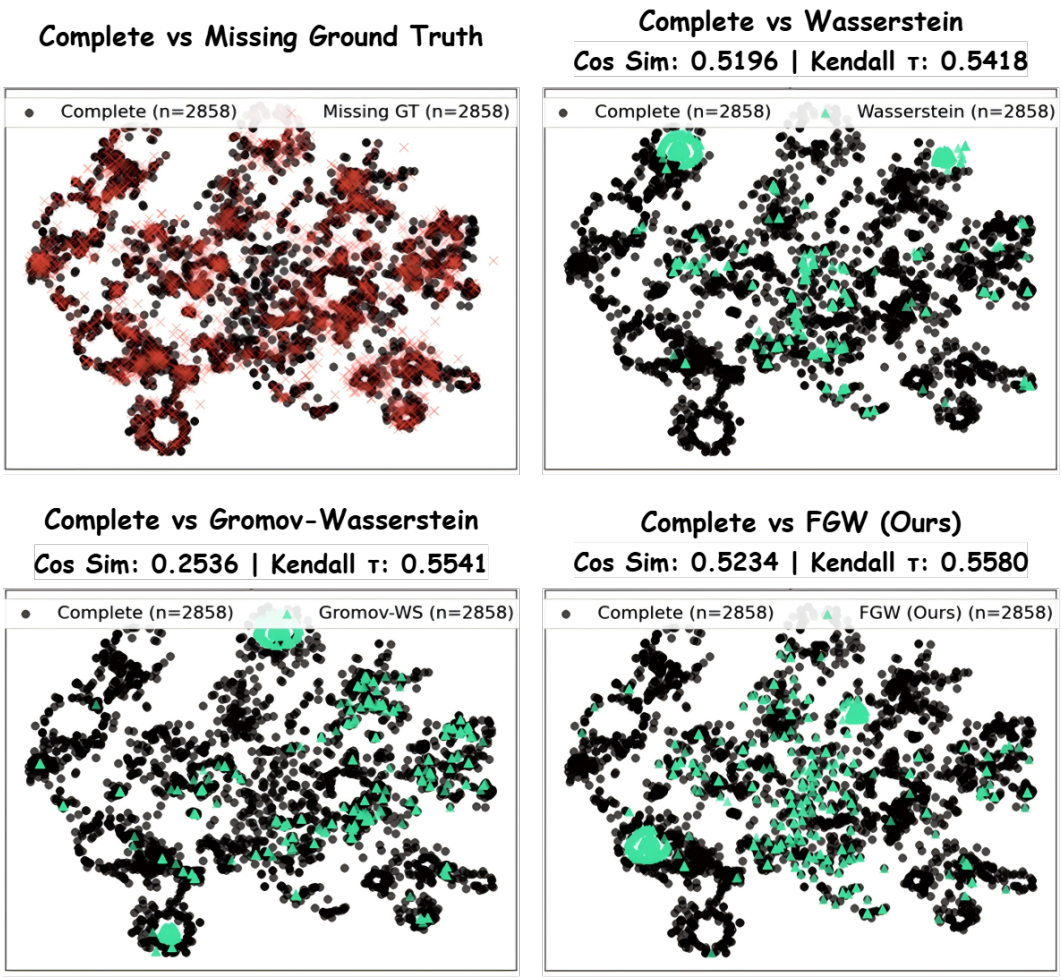

Figure 4: **t-SNE Comparison of Completed Hidden Representations.** t-SNE comparison of completed hidden representations under the missing-image-and-mask setting with a 50% availability ratio, evaluated by cosine similarity and Kendall's τ. FGW achieves a favorable balance between feature-level fidelity and structural preservation.

Table 4 compares OT-MIMIC against image-only baselines that are trained on the full unimodal dataset. This is not an apples-to-apples comparison since OT-MIMIC additionally consumes text at training and test time, and the LViT-derived text descriptions encode coarse spatial location and laterality (Appendix E). The intent of the table is therefore not to claim that FGW modality completion alone surpasses image-only segmentation, but to quantify the value of leveraging text, even after modality completion, when paired image supervision is scarce.

E Experimental Details

Evaluation Metrics. To quantitatively assess segmentation performance, we report two standard overlap-based metrics: the Dice Coefficient and the mean Intersection-over-Union (mIoU) (Jaccard index). Let N denote the total number of test samples. For the i-th sample, let P_i and G_i represent

Model	Modality		Missing Modality	Availability Ratio	QaTa-COV19		MosMedData+	
	Image	Text			Dice	mIoU	Dice	mIoU
U-Net Ronneberger et al. (2015)	√	×	–	–	0.853	0.744	0.716	0.558
U-Net++ Zhou et al. (2018)	√	×	–	–	0.839	0.722	0.714	0.555
nnU-Net Isensee et al. (2021)	√	×	–	–	0.814	0.686	0.688	0.524
Attention U-Net Oktay et al. (2018)	√	×	–	–	0.803	0.671	0.665	0.499
Trans U-Net Chen et al. (2021)	√	×	–	–	0.826	0.703	0.653	0.485
Swin U-Net Cao et al. (2022)	√	×	–	–	0.798	0.664	0.672	0.506
U-KAN Li et al. (2025)	√	×	–	–	0.829	0.708	0.657	0.489
MedSAM Ma et al. (2024)	√	×	–	–	0.649	0.481	0.441	0.283
OT-MIMIC	√	√	Text	1%	**0.888**	**0.799**	**0.773**	**0.630**
			Text	5%	**0.895**	**0.809**	**0.768**	**0.623**
			Text	10%	**0.898**	**0.814**	**0.778**	**0.637**
			Text	25%	**0.908**	**0.831**	**0.774**	**0.631**
			Image + mask	25%	**0.903**	**0.823**	**0.752**	**0.603**

Table 4: **Comparison between image-only baselines with fully observed images and OT-MIMIC under limited fully paired multimodal supervision.**

the predicted and ground-truth foreground masks, respectively. These metrics are formulated as:

$$\begin{aligned} \text{Dice} &= \frac{1}{N}\sum_{i=1}^{N}\frac{2|P_i \cap G_i|}{|P_i| + |G_i|}, \\ \text{mIoU} &= \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i \cap G_i|}{|P_i \cup G_i|}. \end{aligned} \tag{12}$$

For probabilistic outputs, binary masks are generated by applying a sigmoid activation followed by a fixed threshold of 0.5.

Datasets and Pre-processing. **QaTa-COV19** Yamac et al. (2021) comprises 9,258 chest X-ray radiographs annotated with pixel-level lesion masks. **MosMedData+** Morozov et al. (2020) consists of 2,729 CT slices of pulmonary infections. Following Li et al. (2023), we use the released textual descriptions, which summarize infection laterality, the number of affected regions, and their approximate spatial locations, and adopt the official splits from Li et al. (2023): QaTa-COV19 uses 5,716 training, 1,429 validation, and 2,113 test samples, while MosMedData+ uses 2,183 training, 273 validation, and 273 test samples. Images and ground-truth masks are resized to 224×224 and channel-wise normalized using dataset-wide mean/std.

Note on text source. The textual descriptions accompanying QaTa-COV19 and MosMedData+ in Li et al. (2023) are derived from the same radiology reports that informed the mask annotations and contain coarse spatial information (laterality, number of affected regions, approximate location). They should therefore be interpreted as informative, label-correlated clinical text rather than independent supervision. This caveat is most salient for Table 4, where the comparison to image-only baselines should not be read as an apples-to-apples "OT-MIMIC beats image-only segmentation"—rather, it quantifies the value of leveraging clinical text (with modality completion when paired records are scarce) on top of an image-only pipeline. The text content is identical across all baselines that consume it (LViT, GuideSeg, ProLearn, OT-MIMIC), so OT-MIMIC does not receive additional text information unavailable to the competing text-aware baselines in Tables 1–2.

Implementation Details. All experiments are implemented in PyTorch and run on a single NVIDIA H100 GPU (80GB HBM3). We use AdamW with initial learning rate 1×10^{-4}, weight decay 1×10^{-2}, batch size 32, and cosine annealing to $\eta_{\min} = 10^{-6}$. Early stopping with a patience of 20 epochs is applied on validation performance, with a max of 50 epochs. We use the open-source BiomedCLIP-PubMedBERT-256-vit_base_patch16_224 checkpoint Zhang et al. (2023) as the foundation model, the BiomedCLIP encoder weights are *frozen* throughout training. The projection module is a 2-layer

MLP with GELU activation, trained jointly with the segmentation head. Image spatial features pass through a ConvNeXt-based U-Net Liu et al. (2022) encoder, completed CLIP embeddings condition the U-Net decoder via the language-guided fusion mechanism of Zhong et al. (2023).

FGW solver and differentiability. The FGW objective uses $\alpha = 0.5$ and entropic regularization $\epsilon = 0.1$, solved per batch with the `ot.gromov.entropic_fused_gromov_wasserstein` routine in POT Flamary et al. (2024) (max 100 outer iterations, Sinkhorn inner tolerance 10^{-9}). The transport plan $\widehat{\mathbf{T}}$ is *detached* from the autograd graph (no gradient through the FGW solver); the projection MLP is supervised through (i) the image–text matching loss on bank-anchor pairs and (ii) the downstream segmentation loss on modality-completed features. Differentiability of $\widehat{T}$ is not required by the theoretical formulation, although the deployed POT solver remains outside the formal guarantee for the reasons discussed in Appendix B.6. We treat $\widehat{\mathbf{T}}$ as a stationary point of the non-convex FGW objective rather than a global optimum.

Marginals and effective neighborhood. We use balanced FGW with uniform marginals. Each query row is normalized to a probability distribution before barycentric completion. The capacity ablation in Table 3 shows that performance remains stable over moderate bank sizes.

Reproducibility Notes. For Table 1, the ProLearn Ye et al. (2025) entries are taken from the original paper because the text-query component is not released in the open-source code. The ProLearn missing-image-and-mask entries in Table 2 are obtained from a controlled rerun of the public checkpoint and code using the same splits and missingness seeds. In both cases, the comparison follows the ProLearn experimental setting used in the corresponding table. For all other methods, including LViT, GuideSeg, OT-MIMIC, the image-only baselines in Table 4, and MedSAM, results are obtained through our own evaluation under the same splits, preprocessing, and missingness masks. For MedSAM evaluation, we keep the model frozen and provide an oracle box prompt for each validation and test image, computed as the tight bounding box enclosing the ground-truth mask. To ensure a leakage-free evaluation on QaTa-COV19, which is part of MedSAM's training set, we initialize MedSAM from the vanilla SAM backbone with `sam_vit_b_01ec64.pth`. For MosMedData+, the official `medsam_vit_b.pth` weights are utilized.

F Ablation Study

Supplementary Table 5 evaluates the sensitivity of OT-MIMIC to the memory bank capacity K under the missing-text setting across both datasets. Capacity is reported as a percentage of the total training set size, with the absolute number of stored samples provided in parentheses. We observe a consistent trend with missing-image-and-mask scenarios: performance quickly saturates once K reaches a moderate range (e.g., 5%–25%), and further increasing the capacity yields only marginal gains. These results show that OT-MIMIC remains stable across moderate bank capacities and does not require a large fraction of the training set to be stored as anchors.

G Additional Controlled, Cross-Domain, and Diagnostic Experiments

This section provides additional controlled, cross-domain, and diagnostic experiments. Unless otherwise stated, each entry reports test Dice and mIoU and uses the same data split, missingness mask, backbone, memory-bank construction, and training schedule as the corresponding OT-MIMIC experiment.

G.1 Fixed-pipeline ablation of the completion rule

Table 6 isolates the completion rule while holding the rest of the pipeline fixed. To separate the benefit of retaining incomplete records from the benefit of FGW itself, we change only the completion rule. "Nearest anchor" copies the bank entry with the closest observed representation, while "feature mean" uses the mean representation of the bank anchors. FGW consistently improves downstream

Dataset	Capacity	50%		25%		10%		5%		1%	
		Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
QaTa-COV19	1% (57)	**0.911**	**0.837**	**0.908**	0.831	0.897	0.813	**0.895**	**0.811**	0.884	0.792
	5% (286)	0.908	0.831	**0.908**	**0.832**	0.896	0.811	**0.895**	**0.811**	**0.888**	**0.799**
	10% (572)	0.908	0.831	0.906	0.829	0.897	**0.814**	**0.895**	**0.811**	0.886	0.796
	25% (1429)	0.908	0.831	**0.908**	0.831	**0.898**	**0.814**	**0.895**	0.809	**0.888**	**0.799**
MosMed Data+	1% (22)	**0.777**	**0.635**	0.773	0.629	0.773	0.630	**0.772**	**0.629**	**0.773**	**0.630**
	5% (109)	0.774	0.632	0.771	0.628	**0.778**	**0.637**	0.770	0.627	**0.773**	**0.630**
	10% (218)	0.773	0.630	0.772	0.629	0.775	0.633	0.770	0.626	0.772	0.629
	25% (546)	0.775	0.633	**0.774**	**0.631**	**0.778**	**0.637**	0.770	0.626	**0.773**	**0.630**

Table 5: **Ablation on memory bank capacity under missing-text scenarios on QaTa-COV19 and MosMedData+.** We report Dice and mIoU at text availability, or complete-case, ratios of 50%, 25%, 10%, 5%, and 1%. Memory bank capacity is reported as a fraction of the full training set, with the corresponding number of stored samples in parentheses.

performance in all ten QaTa-COV19 cells, with the largest margin under severe missing-image-and-mask conditions.

Missing setting	Completion	50%		25%		10%		5%		1%	
		Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
Text	Nearest anchor	0.896	0.812	0.887	0.797	0.883	0.790	0.879	0.784	0.877	0.781
	Feature mean	0.897	0.813	0.891	0.803	0.887	0.798	0.879	0.783	0.875	0.778
	FGW	**0.908**	**0.831**	**0.908**	**0.831**	**0.898**	**0.814**	**0.895**	**0.809**	**0.888**	**0.799**
Image + mask	Nearest anchor	0.901	0.819	0.893	0.806	0.851	0.741	0.821	0.696	0.720	0.562
	Feature mean	0.895	0.810	0.888	0.799	0.872	0.773	0.858	0.752	0.756	0.608
	FGW	**0.907**	**0.830**	**0.903**	**0.823**	**0.879**	**0.785**	**0.860**	**0.755**	**0.800**	**0.667**

Table 6: **Controlled completion-rule ablation on QaTa-COV19.** Only the completion rule changes; all incomplete records and all other training components are fixed.

This comparison directly addresses whether the gain can be explained by simply retaining the incomplete records. Conventional image self-training is not directly defined for either main regime: missing-text records already retain ground-truth masks, whereas missing-image-and-mask records are text-only and contain no image from which a segmentation network can generate a pseudo-mask. A compatible mask-only-missing comparison is therefore reported in Appendix G.4.

G.2 Theory-compatible neural entropic FGW estimator

Table 7 compares the deployed POT solver with a theory-compatible neural estimator. Following Wang & Goldfeld (2026), we amortize entropic transport-plan estimation with a neural parametric model. Given an observed query embedding and a frozen memory bank, the network outputs normalized transport weights; the same barycentric completion rule is then applied. We train the estimator with the entropic FGW objective rather than a pure OT/GW objective. Under the one-epoch double-sample-split assumptions of Theorem 2, this bounded-complexity parametric completion map is theory-compatible. The original POT implementation and the evolving multi-epoch procedure remain outside the theorem.

The neural estimator is competitive in many settings, but POT is generally stronger and remains our empirical default. We therefore distinguish the *theory-compatible* neural realization from the *empirically preferred* POT realization throughout the paper.

G.3 Recent VLM baseline and cross-domain evaluation

We additionally compare with MedCLIPSeg Koleilat et al. (2026), a recent medical VLM with public code, in the missing-image-and-mask regime. We also extend the evaluation to Kvasir-SEG Jha

Dataset	Missing	Solver	50%		25%		10%		5%		1%	
			Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
QaTa-COV19	Text	Neural	0.887	0.796	0.878	0.783	0.878	0.783	0.876	0.780	0.874	0.775
		POT	**0.908**	**0.831**	**0.908**	**0.831**	**0.898**	**0.814**	**0.895**	**0.809**	**0.888**	**0.799**
	Image + mask	Neural	0.894	0.808	0.884	0.793	0.868	0.767	0.841	0.725	0.760	0.613
		POT	**0.907**	**0.830**	**0.903**	**0.823**	**0.879**	**0.785**	**0.860**	**0.755**	**0.800**	**0.667**
MosMedData+	Text	Neural	0.758	0.611	0.768	0.623	0.770	0.626	0.757	0.610	**0.775**	**0.632**
		POT	**0.775**	**0.633**	**0.774**	**0.631**	**0.778**	**0.637**	**0.770**	**0.626**	0.773	0.630
	Image + mask	Neural	0.751	0.602	0.744	0.593	0.708	0.549	0.683	0.518	0.365	0.223
		POT	**0.769**	**0.624**	**0.752**	**0.603**	**0.711**	**0.552**	**0.684**	**0.519**	**0.566**	**0.395**

Table 7: **POT versus the theory-compatible neural FGW estimator.** Results use the same fixed-seed protocol as the main experiments.

et al. (2020), a non-thoracic endoscopic polyp-segmentation dataset. For Kvasir-SEG, we follow the established LViT-style text-construction protocol and use the dataset as a non-thoracic cross-domain stress test.

As shown in Table 8, OT-MIMIC outperforms MedCLIPSeg across all 15 dataset-ratio comparisons.

Dataset	Method	50%		25%		10%		5%		1%	
		Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
QaTa-COV19	MedCLIPSeg	0.892	0.804	0.885	0.793	0.858	0.751	0.842	0.727	0.788	0.650
	OT-MIMIC	**0.907**	**0.830**	**0.903**	**0.823**	**0.879**	**0.785**	**0.860**	**0.755**	**0.800**	**0.667**
MosMedData+	MedCLIPSeg	0.659	0.491	0.626	0.455	0.587	0.415	0.519	0.350	0.206	0.115
	OT-MIMIC	**0.769**	**0.624**	**0.752**	**0.603**	**0.711**	**0.552**	**0.684**	**0.519**	**0.566**	**0.395**
Kvasir-SEG	MedCLIPSeg	0.858	0.751	0.828	0.706	0.779	0.638	0.750	0.600	0.607	0.436
	OT-MIMIC	**0.910**	**0.836**	**0.889**	**0.799**	**0.850**	**0.739**	**0.805**	**0.674**	**0.700**	**0.538**

Table 8: **Recent-VLM comparison under the missing-image-and-mask setting and a Kvasir-SEG cross-domain extension.**

G.4 Auxiliary mask-only semi-supervised comparison

Conventional semi-supervised segmentation assumes that the unlabeled record still contains an image, so it is not directly compatible with our main missing-image-and-mask regime, in which incomplete records are text-only. We therefore introduce an auxiliary *mask-only-missing* protocol in which all images and text remain available but only 1%–50% of samples retain masks. We compare with DuSSS Pan et al. (2025) and an LViT semi-supervised implementation reconstructed from the paper and released components. All methods use the same labeled subsets, splits, and evaluation protocol.

Table 9 shows that OT-MIMIC achieves the best Dice and mIoU across all 15 dataset-ratio cells under the mask-only-missing protocol.

G.5 Diagnostic for privileged text at 1% availability

Because the benchmark text is derived from the same report stream that informed the masks, it contains label-correlated cues. Table 10 reports a diagnostic in which we repeat the 1% missing-image-and-mask experiment after removing lesion count and anatomical location while retaining only coarse infection status and laterality. This reduces the text-only annotations to two templates on QaTa-COV19 and three on MosMedData+ (including a no-infection template).

Fine-grained text therefore contributes to the strongest numbers, but does not account for the full margin: with weakened text, Dice remains 0.783 on QaTa-COV19 versus 0.732/0.667/0.676 for GuideSeg/LViT/ProLearn, and 0.557 on MosMedData+ versus 0.512/0.316/0.150. Removing lesion-count and location cues causes only modest degradation. OT-MIMIC still retains Dice margins of 5.1

Dataset	Method	50%		25%		10%		5%		1%	
		Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
QaTa-COV19	LViT	0.830	0.709	0.800	0.666	0.801	0.668	0.820	0.694	0.780	0.640
	DuSSS	0.810	0.681	0.812	0.683	0.785	0.646	0.744	0.592	0.582	0.410
	OT-MIMIC	**0.903**	**0.824**	**0.895**	**0.810**	**0.887**	**0.796**	**0.872**	**0.773**	**0.823**	**0.699**
MosMedData+	LViT	0.609	0.438	0.710	0.551	0.685	0.521	0.672	0.506	0.306	0.180
	DuSSS	0.527	0.357	0.567	0.396	0.565	0.394	0.114	0.061	0.049	0.025
	OT-MIMIC	**0.767**	**0.622**	**0.759**	**0.612**	**0.736**	**0.582**	**0.699**	**0.537**	**0.640**	**0.471**
Kvasir-SEG	LViT	0.743	0.591	0.714	0.555	0.678	0.512	0.616	0.445	0.402	0.251
	DuSSS	0.577	0.406	0.578	0.406	0.499	0.332	0.427	0.272	0.274	0.159
	OT-MIMIC	**0.898**	**0.814**	**0.894**	**0.809**	**0.849**	**0.737**	**0.835**	**0.716**	**0.650**	**0.482**

Table 9: **Auxiliary mask-only-missing comparison.** Percentages denote the fraction of training samples retaining segmentation masks; all images and text remain available.

Dataset	Original text	Weakened text	Change
QaTa-COV19	0.800 / 0.667	0.783 / 0.644	$-0.017/-0.023$
MosMedData+	0.566 / 0.395	0.557 / 0.386	$-0.010/-0.010$

Table 10: **Weakened-text diagnostic at 1% image + mask availability.** Entries report Dice and mIoU.

points on QaTa-COV19 and 4.5 points on MosMedData+ over the strongest baseline, showing that fine-grained localization cues are helpful but do not explain the observed gains.

G.6 Computational cost

Table 11 profiles all methods on the same H100 80GB GPU with batch size 8 on QaTa-COV19. OT-MIMIC has the lowest FLOPs among the compared methods in both training and inference, its training latency is close to GuideSeg, and its inference latency is second lowest.

Method	Stage	FLOPs (G)	Latency (ms/iter)	Peak MB
LViT	Train	1300.1	**39.1**	3932.9
	Inference	433.9	**13.5**	1682.6
GuideSeg	Train	873.2	47.2	**2232.8**
	Inference	183.8	19.7	**1352.9**
ProLearn	Train	979.8	89.9	3960.9
	Inference	191.3	46.1	3067.1
OT-MIMIC	Train	**839.8**	49.7	2691.5
	Inference	**144.7**	15.6	1777.3

Table 11: **Computational profile on QaTa-COV19** (H100 80GB, batch size 8).

H Limitations

Data scope and missingness. Our main experiments use two thoracic benchmarks with synthetic missing-completely-at-random (MCAR) masks. Appendix G.3 extends the evaluation to Kvasir-SEG as a non-thoracic stress test, but broader validation on naturally incomplete clinical datasets remains important. Real-world missingness may depend on disease severity, institution, scanner, or clinical workflow and can therefore differ from the controlled missingness considered here.

Text provenance. QaTa-COV19 and MosMedData+ use LViT-derived text annotations containing coarse laterality, lesion-count, and location cues. These annotations are correlated with segmentation

labels and should be viewed as informative clinical context rather than independent supervision. Appendix G.5 removes lesion-count and location cues at 1% availability and shows only a modest performance decrease, suggesting that fine-grained text contributes to, but does not fully explain, the observed gains.

Empirical and theoretical scope. Our controlled experiments isolate the completion rule and compare Wasserstein, GW, and FGW at the representation level. Additional downstream transport comparisons and repeated runs would further strengthen fine-grained empirical conclusions. Theorem 1 analyzes a bounded-complexity parametric completion map under one-epoch sample splitting. The neural estimator in Appendix G.2 satisfies this setting under the stated assumptions, whereas the empirically preferred multi-epoch POT implementation remains outside the formal guarantee.

I BROADER IMPACTS.

This work aims to improve medical image segmentation under incomplete multimodal supervision, which may have positive societal impact by enabling better use of fragmented clinical datasets and reducing reliance on fully paired annotations. In settings where annotations or modalities are scarce, such methods could help make training pipelines more data-efficient and broaden the utility of existing medical archives.

More broadly, the proposed framework may help lower barriers to developing medical AI systems in realistic healthcare environments, where imaging, text, and segmentation labels are often incomplete or unevenly available. By making better use of partially observed samples, such methods may support retrospective research and model development in resource-constrained settings, where collecting large-scale fully annotated multimodal datasets is especially difficult. We view this work as a research step toward more robust and practical multimodal learning for medical data.

J MORE VISUALIZATION

We complement the representative panel of Figure 3 (main paper) with additional qualitative comparisons under severe missing-modality settings. Figure 5 presents additional cases under the missing-image-and-mask setting; Figure 6 shows additional examples under the missing-text setting. Consistent with the main paper, OT-MIMIC generally produces more coherent masks with fewer spurious regions and better boundary adherence than prior baselines, especially in challenging cases with fragmented or low-contrast lesions.

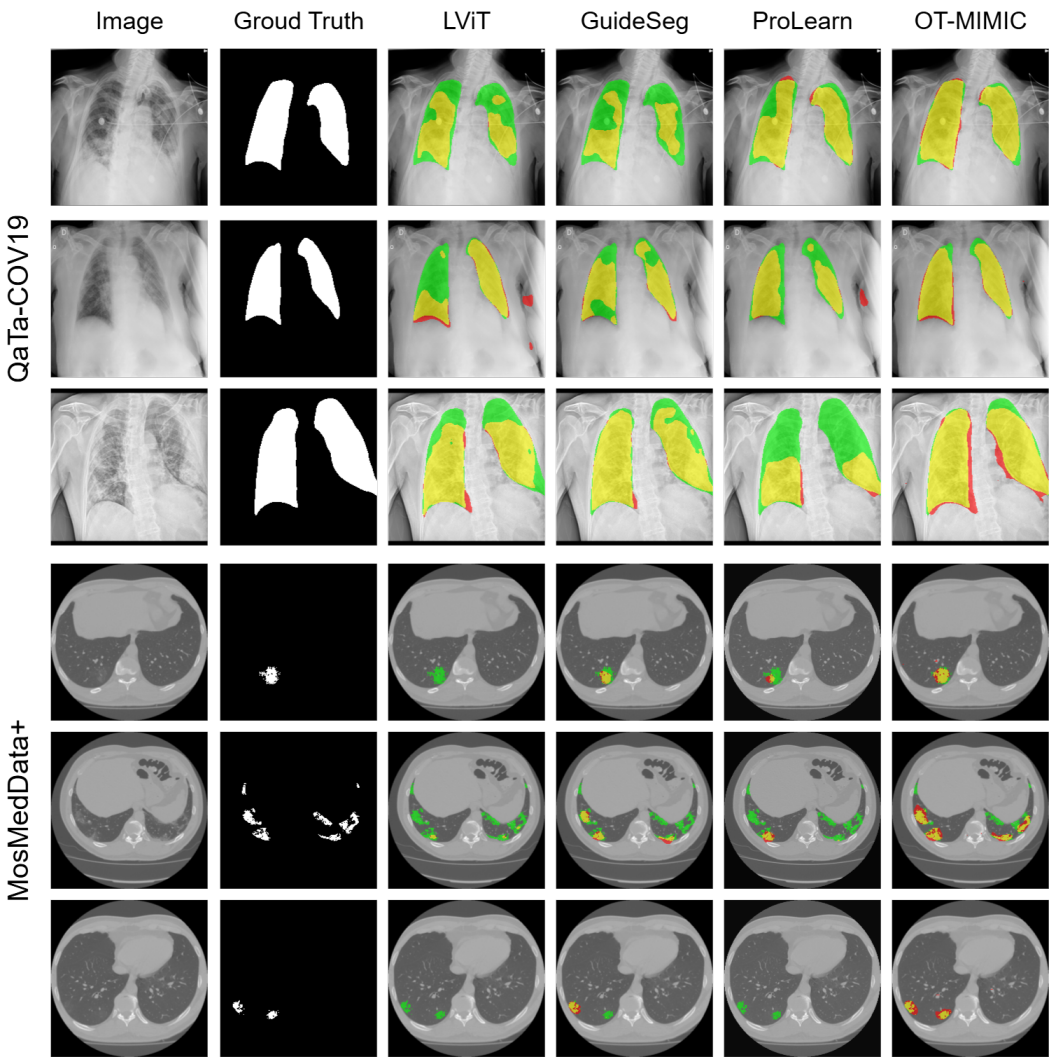

Figure 5: **Qualitative comparisons under the missing-image-and-mask setting.** Visualizations are provided for the QaTa-COV19 and MosMedData+ benchmarks. **Green**, **red**, and **yellow** regions represent the ground truth (GT), model predictions, and their accurate overlap, respectively.

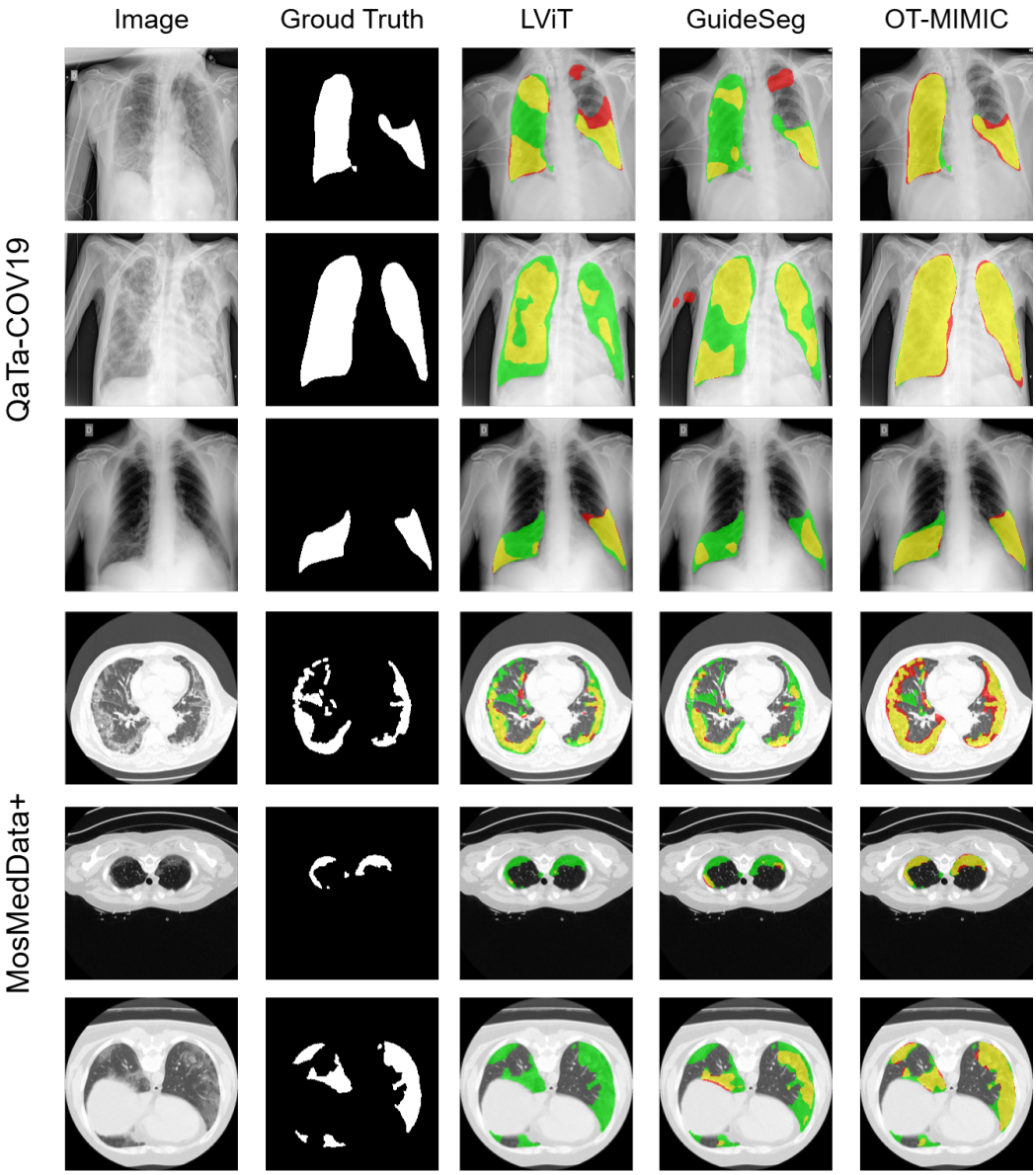

Figure 6: **Qualitative comparisons under the missing-text setting.** Visualizations are provided for the QaTa-COV19 and MosMedData+ benchmarks. **Green**, **red**, and **yellow** regions represent the ground truth (GT), model predictions, and their accurate overlap, respectively.