ALIGN WHAT IS OBSERVED: AVAILABILITY-CONSTRAINED UNBALANCED OPTIMAL TRANSPORT FOR MULTIMODAL SURVIVAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models often assume that every input token is observed and should participate in cross-modal alignment. This assumption is overly strict in light of clinical data with multiple modalities, where some clinical variables may be absent, invalid, or only weakly related to a particular image region. Balanced optimal transport (OT) compounds the problem by forcing every token to send and receive prescribed mass, whereas conventional unbalanced OT (UOT) relaxes the marginals but does not by itself exclude unavailable observations. In light of the limitation, we introduce MASKED-UOT, a differentiable alignment module that combines an availability-constrained transport support, patient-specific learned marginal priors, and KL-relaxed mass conservation. The resulting plan is computed with generalized Sinkhorn iterations and used to transfer information bidirectionally between local 3D CT tokens and structured clinical tokens. A marginal-prior regularizer prevents degenerate concentration, while an optional patient-level contrastive objective complements the local transport loss. In a retrospective study of 429 CT sessions from 330 patients, MASKED-UOT with contrastive learning achieved an AUC of 0.7176, outperforming its balanced-OT counterpart and single-modality models using the same backbone.

1 INTRODUCTION

Multimodal learning (Ektefaie et al., 2023; Reza et al., 2024; Yuan et al., 2025) seeks to combine complementary evidence from multiple types of data, referred to as modalities, such as text, audio, images, or video. The primary objective of multimodal learning is to construct a unified embedding space where semantically related data from different modalities can be aligned, enabling complex tasks such as cross-modal retrieval (Wang et al., 2025), autonomous systems (Lekha et al., 2026), and scientific discovery (Lopez et al., 2025). Its success depends not only on strong unimodal encoders, but also on an alignment mechanism that determines which elements across modalities should interact. This issue is particularly important in clinical prediction (Swanson et al., 2023) and radiation therapy practice (Huynh et al., 2020). For example, a planning CT volume contains hundreds of local anatomical representations, whereas a structured clinical record may contain only a small set of heterogeneous variables (e.g., age, gender, race, TNM stage, histology, medical history, and target volume). Their relationship is neither one-to-one nor equally reliable for every patient.

Optimal transport (OT) (Lee et al., 2019) offers a principled way to represent such soft correspondences. Given two empirical measures, OT is to find a coupling that minimizes the global cost of moving mass between them (You & Liu, 2025). Entropic regularization (Lin et al., 2022) makes the coupling differentiable and computationally practical through Sinkhorn scaling (Cuturi, 2013). OT has therefore become an attractive alternative to concatenation or unconstrained cross-attention for multimodal alignment (Zhu et al., 2026). However, balanced OT enforces exact row and column marginals. In incomplete or noisy data, this constraint can allocate mass to an unavailable clinical variable or force a weakly related token to participate in cross-modality alignment.

Unbalanced OT (UOT) relaxes exact mass conservation with divergence penalties (Chizat et al., 2018; Séjourné et al., 2019). It can suppress unmatched mass and is consequently better suited to partial overlap. Yet two limitations remain when using UOT in multimodal learning: (i) marginal

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.

relaxation does not encode whether a token is actually observed: an invalid entry needs to be removed from the admissible support, not merely assigned a favorable or unfavorable cost; and (ii) fixed uniform marginals assume that all available tokens should begin with equal importance. This is inappropriate when the relevance of image regions and clinical variables varies across patients.

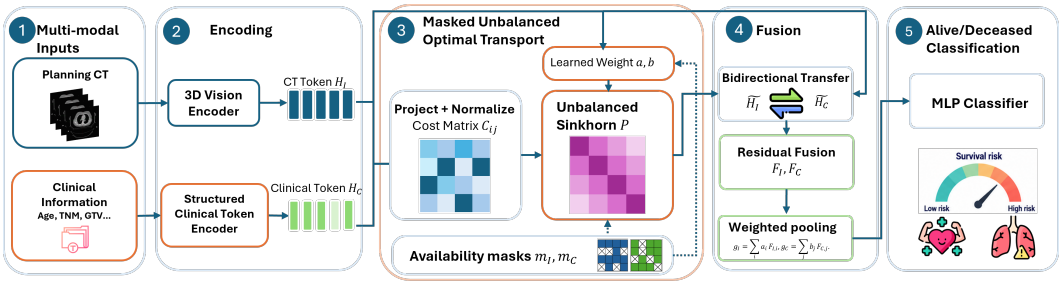


Figure 1: Overview of the proposed MASKED-UOT. Binary masks remove unavailable tokens from the admissible transport support. Patient-specific marginal priors assign initial importance only to available tokens, while KL-relaxed UOT determines the mass that is actually transported. The coupling transfers information in both directions before patient-level prediction.

We address these difficulties with MASKED-UOT, illustrated in Figure 1. The method separates three concepts that are often conflated: *availability* determines whether a pair is allowed to receive transport mass; *marginal priors* encode patient-specific preferences over the available tokens; and *unbalanced transport* determines how much of that preferred mass should actually be matched. A binary support mask enforces availability exactly. Lightweight scoring networks produce masked-softmax marginal priors, and KL penalties allow the optimized row and column masses to depart from those priors. The resulting plan provides interpretable local correspondences and enables bidirectional cross-modal feature transfer, potentially enhancing the robustness of multimodal learning.

The proposed formulation is evaluated in a concrete but challenging setting: binary overall-survival prediction from 3D planning CT and structured clinical variables. The study includes a patient-disjoint institutional cohort and the public NSCLC-Radiomics (LUNG1) cohort (Aerts et al., 2014). We compare balanced OT, UOT with availability constraints, and their patient-level contrastive variants using matched encoders. The results reveal a more nuanced picture than a blanket superiority claim. Availability-aware UOT improves ranking performance on the institutional cohort and benefits a MedicalNet encoder (Chen et al., 2019) on LUNG1. Our contributions are as follows:

- We formulate availability-constrained MASKED-UOT for multimodal alignment, explicitly separating hard observation support from soft transport mass. This prevents missing or invalid tokens from entering the coupling.
- We introduce patient-specific marginal priors over the available tokens and regularize them toward availability-conditioned uniform measures. MASKED-UOT can deviate from these priors when the evidence supports partial matching.
- We derive a differentiable generalized Sinkhorn solver and a mass-normalized bidirectional feature-transfer rule suitable for unequal token counts and unequal transported mass.
- We provide a two-cohort clinical case study with matched balanced-OT controls, encoder comparisons, and metric-level analysis under moderate and severe alive/deceased class imbalance.

2 RELATED WORK

Optimal transport for representation alignment. OT aligns heterogeneous distributions through globally coupled correspondences, while entropic regularization enables efficient and differentiable Sinkhorn optimization (Cuturi, 2013). UOT relaxes exact marginal constraints and supports distributions with unequal mass or partial overlap (Chizat et al., 2018; Séjourné et al., 2019). These

properties have motivated OT-based multimodal alignment and missing-modality inference; for example, MM-Align learns alignment dynamics for incomplete modality sequences (Han et al., 2022). In contrast, MASKED-UOT addresses unequal, non-temporal CT and clinical token sets by restricting transport to available observations, learning patient-specific marginal priors, and normalizing feature transfer using optimized UOT marginals.

Learning with incomplete modalities. Incomplete multimodal learning commonly employs feature reconstruction (Ma et al., 2021), cross-modal generation (Pan et al., 2025), knowledge distillation (Wang et al., 2020), modality dropout (Gu et al., 2025), or shared-specific representation learning (Wang et al., 2023). Although these strategies can recover complementary information, the reconstruction process may introduce content lacking a factual basis, whereas distillation methods typically require representative and fully observed training samples. MASKED-UOT instead excludes missing observations from the admissible transport support and uses UOT to suppress noisy, irrelevant, or unmatched features. This approach preserves data provenance and avoids assigning artificial certainty to unobserved clinical variables.

Multimodal survival prediction. Deep learning has advanced cancer prognosis through convolutional neural networks (O'shea & Nash, 2015), Vision Transformers (Khan et al., 2022), and three-dimensional image encoders (Ferretti & Corino, 2025). Multimodal large language models (Yin et al., 2024) further integrate imaging and clinical text through Chain-of-Thought reasoning (Wei et al., 2022), semantic guidance, and cross-modal attention (Kim et al., 2024). CLIP-based alignment (Radford et al., 2021), XMedGPT (Yang et al., 2025), KEMM (Zhao et al., 2025), SurvAgent (Huang et al., 2025), and the LungCURE benchmark (Hao et al., 2026) demonstrate the potential of reasoning-enhanced medical multimodal models. Meanwhile, radiomic and deep CT representations remain important for lung-cancer prognosis (Aerts et al., 2014), with MC3-18 supporting efficient volumetric feature extraction (Tran et al., 2018) and MedicalNet providing medical-domain 3D pretraining (Chen et al., 2019). Unlike global feature concatenation or free-form language reasoning, MASKED-UOT aligns local CT regions with observed clinical-variable tokens, producing a patient-specific fusion operator and an analyzable cross-modal correspondence map.

3 PROPOSED METHOD

3.1 PROBLEM STATEMENT

For patient p, let $X_I^{(p)}$ denote a 3D CT volume and $X_C^{(p)} = (x_{C,1}^{(p)}, \ldots, x_{C,N_C}^{(p)})$ denote structured clinical variables. We perform binary vital-status classification, where $y_p = 1$ indicates recorded deceased status and $y_p = 0$ indicates alive status. Modality-specific encoders produce token matrices:

$$H_I = f_I(X_I) \in \mathbb{R}^{N_I \times d_I}, \qquad H_C = f_C(X_C) \in \mathbb{R}^{N_C \times d_C}. \tag{1}$$

The CT encoder retains a spatial grid of local features, while the clinical encoder produces one token for each of nine variables: age, sex, smoking, location, stage, TNM, histology, gross tumor volume (GTV), and planning target volume (PTV). Continuous variables are encoded by variable-specific MLPs receiving normalized values and availability indicators. Categorical values are mapped to deterministic hash-bucket indices and encoded using variable-specific embedding tables, with index zero reserved for missing values. A learned variable-identity embedding is added to each clinical token, followed by layer normalization.

Binary masks $m_I \in \{0,1\}^{N_I}$ and $m_C \in \{0,1\}^{N_C}$ indicate token availability. For CT, the default masking procedure thresholds the preprocessed volume before intensity augmentation, applies the same geometric transformations to the voxel mask, and uses adaptive max pooling to map it onto the CT token grid. The default intensity threshold is zero, and pooled values greater than 0.5 indicate valid tokens. This mask represents CT support, not tumor segmentation or learned anatomical relevance. Clinical availability is determined field-wise; the TNM token is available only when all three components are present.

Each sample must contain at least one valid CT token and one observed clinical token:

$$\mathbf{1}^\top m_I > 0, \qquad \mathbf{1}^\top m_C > 0. \tag{2}$$

Individual missing clinical fields are excluded, but entirely missing modalities are unsupported. Unavailable token features are zeroed before alignment. We omit the patient index unless needed and use $\mathbf{1}$ for an all-ones vector of the appropriate dimension, LN for layer normalization, and $[\cdot;\cdot]$ for feature concatenation.

Projection heads map both modalities into a shared d-dimensional space:

$$Z_I = H_I W_I^\top, \qquad Z_C = H_C W_C^\top, \tag{3}$$

where $W_I \in \mathbb{R}^{d\times d_I}$ and $W_C \in \mathbb{R}^{d\times d_C}$. We normalize each projected token as

$$\widehat{z}_{I,i} = \frac{z_{I,i}}{\max(\|z_{I,i}\|_2, \delta_z)}, \qquad \widehat{z}_{C,j} = \frac{z_{C,j}}{\max(\|z_{C,j}\|_2, \delta_z)}, \tag{4}$$

where $\delta_z > 0$ is a numerical normalization constant. The cross-modal cosine-distance cost matrix is

$$C_{ij} = 1 - \langle \widehat{z}_{I,i}, \widehat{z}_{C,j} \rangle, \qquad C \in \mathbb{R}^{N_I \times N_C}, \tag{5}$$

where i and j index CT and clinical tokens, respectively.

3.2 Availability support and learned marginal priors

We define the active token sets

$$\mathcal{I} = \{i : m_{I,i} = 1\}, \qquad \mathcal{C} = \{j : m_{C,j} = 1\}, \tag{6}$$

and the admissible transport support

$$S_{ij} = m_{I,i} m_{C,j}, \quad \mathcal{P}(S) = \left\{ P \in \mathbb{R}_+^{N_I \times N_C} : P_{ij} = 0 \text{ if } S_{ij} = 0 \right\}. \tag{7}$$

An unavailable observation therefore cannot receive transport mass, regardless of its encoded value or pairwise cost.

Scalar scoring functions s_I and s_C produce patient-specific marginal priors:

$$a_i = \frac{m_{I,i} \exp(s_I(h_{I,i}))}{\sum_k m_{I,k} \exp(s_I(h_{I,k}))}, \qquad b_j = \frac{m_{C,j} \exp(s_C(h_{C,j}))}{\sum_\ell m_{C,\ell} \exp(s_C(h_{C,\ell}))}. \tag{8}$$

The scoring functions are learned affine maps, initialized to zero so that the initial priors are uniform over available tokens. These priors specify preferences for allocating transport mass; they are not calibrated clinical importance scores.

For nonnegative vectors, we use the generalized KL divergence (Cui et al., 2026)

$$\mathrm{KL}(r\|q) = \sum_k \left[r_k \log \frac{r_k}{q_k} - r_k + q_k \right], \tag{9}$$

If $r_k = 0$ and $q_k > 0$, the logarithmic term is defined as zero, so the contribution is q_k. If both values are zero, the contribution is zero. If $r_k > 0$ but $q_k = 0$, the contribution is infinite. In our masked objectives, we sum only over available tokens and exclude missing or invalid tokens.

To discourage premature concentration of the learned priors, we define

$$\mathcal{R}_{\text{prior}} = \mathrm{KL}(a\|u_I) + \mathrm{KL}(b\|u_C), \tag{10}$$

where $u_I = \frac{m_I}{\mathbf{1}^\top m_I}$ and $u_C = \frac{m_C}{\mathbf{1}^\top m_C}$.

3.3 Availability-constrained unbalanced transport

Given C, a, and b, we define the entropic unbalanced transport objective

$$\mathcal{J}_{\mathrm{UOT}}(P) = \langle P, C \rangle + \epsilon \sum_{i\in\mathcal{I},\, j\in\mathcal{C}} P_{ij}(\log P_{ij} - 1) + \tau_I \,\mathrm{KL}(P\mathbf{1}\|a) + \tau_C \,\mathrm{KL}(P^\top \mathbf{1}\|b), \tag{11}$$

and its mathematical minimizer

$$P^* \in \arg\min_{P\in\mathcal{P}(S)} \mathcal{J}_{\mathrm{UOT}}(P), \qquad \epsilon, \tau_I, \tau_C > 0. \tag{12}$$

Here, $\langle P, C \rangle = \sum_{i,j} P_{ij} C_{ij}$. The marginal KL penalties allow transported masses to deviate from the learned priors. Consequently, neither $P\mathbf{1} = a$ nor $P^\top \mathbf{1} = b$ is imposed, and the total transported mass need not equal one.

Log-domain generalized Sinkhorn solver. We approximate Equation (12) using generalized Sinkhorn scaling (Chizat et al., 2018). On the active support, let $L_{ij} = -C_{ij}/\epsilon$ and define

$$\rho_I = \frac{\tau_I}{\tau_I + \epsilon}, \qquad \rho_C = \frac{\tau_C}{\tau_C + \epsilon}. \tag{13}$$

The positive scaling vectors have dimensions $u \in \mathbb{R}^{|\mathcal{I}|}_{>0}$ and $v \in \mathbb{R}^{|\mathcal{C}|}_{>0}$. Writing $\alpha = \log u$ and $\gamma = \log v$, initialize $\alpha^{(0)} = 0$ and $\gamma^{(0)} = 0$ and iterate

$$\alpha_i^{(t+1)} = \rho_I \left[\log a_i - \mathrm{LSE}_{j\in\mathcal{C}}\left(L_{ij} + \gamma_j^{(t)}\right)\right], \tag{14}$$

$$\gamma_j^{(t+1)} = \rho_C \left[\log b_j - \mathrm{LSE}_{i\in\mathcal{I}}\left(L_{ij} + \alpha_i^{(t+1)}\right)\right], \tag{15}$$

where $\mathrm{LSE}_k(x_k) = \log \sum_k \exp(x_k)$ denotes the log-sum-exp operator. In numerical evaluation, positive quantities are floored before taking logarithms.

After T iterations, the computed plan is

$$P_{ij}^{(T)} = \begin{cases} \exp\left(L_{ij} + \alpha_i^{(T)} + \gamma_j^{(T)}\right), & S_{ij} = 1, \\ 0, & S_{ij} = 0. \end{cases} \tag{16}$$

We use $T = 50$ by default and distinguish the computed plan $P^{(T)} \approx P^*$ from the exact minimizer. The solver is unrolled during training, allowing gradients to propagate to the trainable encoders, projection heads, and marginal-scoring functions.

Relation to standard transport formulations. With all tokens available, Equation (12) is the standard entropic UOT objective with learned marginal priors. For fixed C, a, b, and ϵ, taking $\tau_I, \tau_C \to \infty$ enforces the marginal constraints and recovers balanced entropic OT in the limit.

3.4 Bidirectional transport and fusion

Let $P = P^{(T)}$ denote the final transport plan obtained after T iterations of the UOT optimization. Define its row and column masses:

$$r = P\mathbf{1}, \qquad c = P^\top \mathbf{1}. \tag{17}$$

Because UOT relaxes the marginal constraints, conditional feature transfer is normalized by these computed masses. We transfer the original modality-specific features:

$$\bar{h}_{I,i} = \frac{\sum_j P_{ij} h_{C,j}}{\max(r_i, \delta)}, \qquad \bar{h}_{C,j} = \frac{\sum_i P_{ij} h_{I,i}}{\max(c_j, \delta)}, \tag{18}$$

where $\delta > 0$ is a numerical floor. Here, $\bar{h}_{I,i} \in \mathbb{R}^{d_C}$ is the clinical feature transferred to CT token i, and $\bar{h}_{C,j} \in \mathbb{R}^{d_I}$ is the CT feature transferred to clinical token j.

To modulate transfer by the fraction of prior mass utilized, we compute

$$\eta_{I,i} = \mathrm{clip}\left(\frac{r_i}{\max(a_i, \delta)}, 0, 1\right), \qquad \eta_{C,j} = \mathrm{clip}\left(\frac{c_j}{\max(b_j, \delta)}, 0, 1\right). \tag{19}$$

These factors are determined by the transport plan and marginal priors. We use utilization-modulated residual fusion:

$$f_{I,i} = \mathrm{LN}_I\left(h_{I,i} + \phi_{CI}(\eta_{I,i}\bar{h}_{I,i})\right), \qquad f_{C,j} = \mathrm{LN}_C\left(h_{C,j} + \phi_{IC}(\eta_{C,j}\bar{h}_{C,j})\right), \tag{20}$$

where $\phi_{CI} : \mathbb{R}^{d_C} \to \mathbb{R}^{d_I}$ and $\phi_{IC} : \mathbb{R}^{d_I} \to \mathbb{R}^{d_C}$ are learned affine projections. Utilization is applied before the affine projection; the projection bias can therefore remain nonzero even when the transferred input is zero. Invalid fused tokens are subsequently set to zero.

Prediction readout. The default model pools fused tokens using the learned marginal priors:

$$g_I = \sum_i a_i f_{I,i}, \qquad g_C = \sum_j b_j f_{C,j}, \qquad g = [g_I; g_C]. \tag{21}$$

Thus, the default model shares its alignment priors with its prediction readout. The classifier receives the concatenated modality representations and produces a single logit:

$$\ell = f_{\mathrm{pred}}(g), \qquad \hat{y} = \sigma(\ell). \tag{22}$$

3.5 TRAINING OBJECTIVE

Classification loss. For a minibatch of B samples, the outcome loss is

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{B}\sum_{p=1}^{B}\left[w_{+}y_p \log \hat{y}_p + (1-y_p)\log(1-\hat{y}_p)\right], \tag{23}$$

where w_+ is the positive-class weight. The default setting $w_+ = 1$ gives unweighted binary cross-entropy (BCE). This loss is evaluated directly from logits using a numerically stable BCE-with-logits implementation.

Patient-level contrastive alignment. An optional symmetric InfoNCE loss (Oord et al., 2018) aligns modality-specific patient representations before transport and fusion. For $M \in \{I, C\}$, let

$$\mu_M^{(p)} = \frac{\sum_k m_{M,k}^{(p)} h_{M,k}^{(p)}}{\sum_k m_{M,k}^{(p)}}, \qquad e_M^{(p)} = \mathrm{Normalize}\Big(\psi_M(\mu_M^{(p)})\Big), \tag{24}$$

where ψ_I and ψ_C are separate projection MLPs and $\mathrm{Normalize}$ denotes numerically stabilized ℓ_2 normalization.

Let $\mathcal{Q}_C = \{q_{C,k}\}_{k=1}^{K_Q}$ and $\mathcal{Q}_I = \{q_{I,k}\}_{k=1}^{K_Q}$ contain normalized, detached embeddings from previous minibatches. For temperature $\theta > 0$, define

$$D_{I\to C}^{(p)} = \sum_{q=1}^{B}\exp\left(\frac{\langle e_I^{(p)}, e_C^{(q)}\rangle}{\theta}\right) + \sum_{k=1}^{K_Q}\exp\left(\frac{\langle e_I^{(p)}, q_{C,k}\rangle}{\theta}\right), \tag{25}$$

and

$$\mathcal{L}_{I\to C} = -\frac{1}{B}\sum_{p=1}^{B}\log\frac{\exp\Big(\langle e_I^{(p)}, e_C^{(p)}\rangle/\theta\Big)}{D_{I\to C}^{(p)}}. \tag{26}$$

The reverse loss $\mathcal{L}_{C\to I}$ is defined analogously by exchanging modalities and using $\mathcal{Q}_I$. The symmetric loss is

$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{2}\left(\mathcal{L}_{I\to C} + \mathcal{L}_{C\to I}\right). \tag{27}$$

Each modality uses a first-in, first-out queue with default capacity 256; the default temperature is $\theta = 0.07$. Queued embeddings are treated as negatives without filtering by patient identity. The queue used to evaluate the current loss contains embeddings from preceding minibatches.

Combined objective. We evaluate the transport objective at the computed plan and average over the minibatch:

$$\overline{\mathcal{J}}_{\mathrm{UOT}} = \frac{1}{B}\sum_{p=1}^{B}\mathcal{J}_{\mathrm{UOT}}^{(p)}\Big(P^{(T,p)}\Big), \overline{\mathcal{R}}_{\mathrm{prior}} = \frac{1}{B}\sum_{p=1}^{B}\mathcal{R}_{\mathrm{prior}}^{(p)}. \tag{28}$$

Then our training objective for MASKED-UOT is

$$\mathcal{L}_{overall} = \mathcal{L}_{\mathrm{BCE}} + \lambda_{\mathrm{OT}}\left(\overline{\mathcal{J}}_{\mathrm{UOT}} + \beta_{\mathrm{prior}}\overline{\mathcal{R}}_{\mathrm{prior}}\right) + \lambda_{\mathrm{CL}}\mathcal{L}_{\mathrm{CL}}. \tag{29}$$

When $\lambda_{\mathrm{CL}} = 0$, the contrastive loss is excluded from training. The prior regularizer is weighted by both β_{prior} and λ_{OT}. For example, setting λ_{OT} to 0.01 and β_{prior} to 0.1, respectively, gives a final weight of 0.001 for the prior regularizer in the total loss.

4 EXPERIMENTS

We propose three research questions to evaluate the proposed MASKED-UOT method: (RQ1) Can MASKED-UOT improve survival prediction performance compared to matched BOT models or unimodal models? (RQ2) Can MASKED-UOT maintain stable performance when the availability of clinical variables changes? (RQ3) What specific impact does removing certain components of the objective function of MASKED-UOT have on model performance?

4.1 EXPERIMENTAL SETTINGS

Datasets and evaluation. We evaluated our method on an institutional lung cancer cohort comprising 429 CT simulation sessions from 330 patients and the public NSCLC-Radiomics (LUNG1) cohort, comprising 416 eligible patients after preprocessing (Aerts et al., 2014). The alive/deceased counts were 286/143 institutional sessions and 49/367 LUNG1 patients. The prediction target was recorded binary vital status, rather than survival at a fixed time horizon. Both cohorts used a shared nine-variable clinical schema, with missing or invalid variables excluded from OT alignment through availability masks.

Fixed held-out test partitions contained 57 sessions from 48 institutional patients and 62 LUNG1 patients. The remaining data were divided into four development folds, with each model trained on three folds and validated on the fourth. Institutional partitions were patient-disjoint, keeping repeated sessions from each patient together. Each fold-trained model was evaluated on the same dataset-specific test partition; results are reported as mean $\pm$ standard deviation across the four models. Classification thresholds maximized Youden's index (Fluss et al., 2005) on the corresponding validation partition and were applied unchanged to the test data.

Models and training. We evaluated MC3-18 (Tran et al., 2018) and MedicalNet (Chen et al., 2019) CT backbones with our designed StructuredClinicalTokenEncoder (see Appendix C.4). Image and clinical tokens were projected into a shared 256-dimensional space for transport-based fusion. Hyperparameters were fixed without a search. Models were trained for up to 100 epochs using a learning rate of 10^{-5} and batch size of 2, with validation-AUC-based early stopping. Unless otherwise specified, $\lambda_{\text{OT}} = 0.1$, $\beta_{\text{prior}} = 0.01$, and $\lambda_{\text{CL}} = 0.1$ when CL was enabled. Detailed cohort preparation, preprocessing, architecture, and training settings are provided in Appendix C.

Baselines. We compare MASKED-UOT with CT-only, clinical-only, text-guided attention (TGA) (Kim et al., 2024), and balanced optimal transport (BOT). Our structured TGA implementation combines MC3-18 3D CT encoder, a contextual structured clinical encoder, bidirectional cross-attention, and TransMIL aggregation. BOT shares MASKED-UOT's encoder architectures and clinical availability masks, but uses uniform marginals, mean CT pooling, and masked mean clinical pooling. MASKED-UOT additionally applies CT-token masking and uses learned marginal priors for transport and pooling. TGA uses BCE loss; both transport models add their respective entropic OT objectives, with marginal-relaxation penalties and learned-prior regularization specific to MASKED-UOT. The MASKED-UOT + CL variant uses the same additional CT-clinical contrastive objective.

Table 1: Performance comparison of TGA, Balanced OT (BOT), and Masked Unbalanced OT (MASKED-UOT), with and without contrastive learning (CL), for binary vital-status classification on an institutional dataset. The cohort-level alive-to-deceased ratio is 2 : 1. Individual configurations report mean $\pm$ standard deviation.

Model	Backbone	AUC	AP	BACC	Precision	Recall	F1-score
CI-only	StructuredClinicalTokenEncoder	0.6948 ± 0.0541	0.5874 ± 0.0850	0.6368 ± 0.0539	0.4514 ± 0.0785	0.7222 ± 0.2796	0.5225 ± 0.0838
CT-only	MC3-18	0.6499 ± 0.1004	0.4758 ± 0.1044	0.5759 ± 0.0410	0.4766 ± 0.1157	0.3889 ± 0.2175	0.3892 ± 0.1067
	MedicalNet	0.6866 ± 0.0959	0.5080 ± 0.1227	0.6213 ± 0.0966	0.4336 ± 0.1405	0.5694 ± 0.3696	0.4600 ± 0.2055
TGA	MC3-18 + ContextualStructuredClinicalEncoder	0.6574 ± 0.0243	0.5286 ± 0.0913	0.6079 ± 0.0607	0.4248 ± 0.0884	0.6389 ± 0.0556	0.5046 ± 0.0543
BOT	MC3-18 + StructuredClinicalTokenEncoder	0.6827 ± 0.0373	0.5454 ± 0.0548	0.6378 ± 0.0401	0.4410 ± 0.0322	0.6667 ± 0.1111	0.5285 ± 0.0452
	MedicalNet + StructuredClinicalTokenEncoder	0.6688 ± 0.0257	0.5355 ± 0.0644	0.6410 ± 0.0481	$\mathbf{0.5010} \pm 0.0931$	0.5833 ± 0.2245	0.5065 ± 0.0942
BOT + CL	MC3-18 + StructuredClinicalTokenEncoder	0.6866 ± 0.0555	0.5423 ± 0.0782	$\mathbf{0.6624} \pm 0.0214$	0.4625 ± 0.0382	$\mathbf{0.7222} \pm 0.1200$	$\mathbf{0.5586} \pm 0.0239$
	MedicalNet + StructuredClinicalTokenEncoder	0.6620 ± 0.0344	0.5351 ± 0.0319	0.6100 ± 0.0307	0.4320 ± 0.0503	0.6111 ± 0.1361	0.4960 ± 0.0292
MASKED-UOT	MC3-18 + StructuredClinicalTokenEncoder	0.7041 ± 0.0585	0.6076 ± 0.0697	0.6384 ± 0.0208	0.4727 ± 0.0469	0.5972 ± 0.1461	0.5168 ± 0.0386
	MedicalNet + StructuredClinicalTokenEncoder	0.6720 ± 0.0260	0.5333 ± 0.0623	0.6245 ± 0.0249	0.4646 ± 0.0560	0.5694 ± 0.1145	0.5024 ± 0.0238
MASKED-UOT + CL	MC3-18 + StructuredClinicalTokenEncoder	$\mathbf{0.7176} \pm 0.0408$	$\mathbf{0.6152} \pm 0.0599$	0.6293 ± 0.0224	0.4413 ± 0.0546	0.6944 ± 0.2335	0.5197 ± 0.0434
	MedicalNet + StructuredClinicalTokenEncoder	0.6542 ± 0.0448	0.5376 ± 0.0638	0.6255 ± 0.0613	0.4559 ± 0.0803	0.5972 ± 0.1145	0.5096 ± 0.0662

4.2 OVERALL COMPARISON

Table 1 shows the overall binary vital-status classification performance of TGA, BOT, MASKED-UOT on the institutional lung cancer dataset. Further results regarding the LUNG1 dataset (Aerts et al., 2014) are provided in Appendix E. MASKED-UOT achieves the best ranking performance using the MC3-18 backbone, it outperforms BOT in both AUC and AP, with or without contrastive learning (CL). MASKED-UOT + CL achieves the highest AUC (0.7176) and AP (0.6152), improving over BOT + CL with the same backbone and clinical encoder by 3.10 and 7.29 percentage

points, respectively. Although MASKED-UOT + CL achieves the highest AUC and AP, BOT-based models perform better on several metrics computed from binary predictions. Specifically, MC3-18-based BOT + CL achieves the highest BACC (0.6624), recall (0.7222), and F1-score (0.5586), while MedicalNet-based BOT achieves the highest precision (0.5010). The benefits of contrastive learning also depend on the configuration. For MASKED-UOT with MC3-18, it improves AUC, AP, and recall but reduces BACC and precision. For MASKED-UOT with MedicalNet backbone, adding CL reduces AUC from 0.6720 to 0.6542. Furthermore, the MedicalNet CT-only baseline achieves a higher AUC score than all MedicalNet-based multimodal configurations but the lowest F1-score. This suggests that CT features alone provide competitive ranking performance, while the multimodal configurations achieve a better balance between precision and recall. These results demonstrate that MASKED-UOT improves AUC and AP with the MC3-18 backbone.

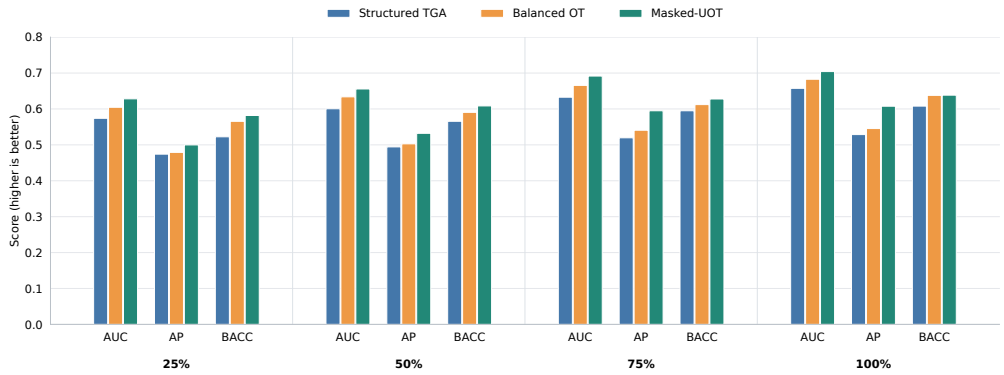


Figure 2: Sensitivity analysis of Structured TGA, Balanced OT, and MASKED-UOT in the face of variations in clinical variable availability from 25% to 100%.

4.3 SENSITIVITY UNDER REDUCED AVAILABILITY.

To evaluate the robustness of MASKED-UOT as the degree of data missingness increases, we investigate how its predictive performance changes as the available clinical information decreases. Information regarding the availability of clinical variables is provided in Figure 3 (Appendix D). This experiment aims to assess whether the advantages of MASKED-UOT persist when additional masking operations further reduce the clinical information used for prediction.

We compare Structured TGA, Balanced OT, and MASKED-UOT at availability levels of 100%, 75%, 50%, and 25%. Availability is controlled by masking originally observed clinical variables during evaluation, while keeping the CT input and trained model weights fixed. We report AUC, average precision (AP), and balanced accuracy (BACC) at each level to assess both performance degradation and the relative ranking of the methods.

Figure 2 shows that all three methods exhibit declining performance as availability decreases from 100% to 25%, indicating that additional missingness makes prediction more challenging. MASKED-UOT consistently achieves the highest AUC, AP, and BACC at every evaluated availability level. At 25% availability, MASKED-UOT achieves an AUC of 0.6281, AP of 0.4998, and BACC of 0.5818, exceeding Balanced OT by 2.35, 2.08, and 1.63 percentage points, respectively. Balanced OT also consistently outperforms Structured TGA across all three metrics and availability levels. These results demonstrate that the performance advantage of MASKED-UOT is maintained regardless of whether the availability of clinical variables is mildly (100% of original clinical variable availability) or severely limited (25% of original clinical variable availability).

4.4 ABLATION STUDY

We conducted a loss-component ablation study to evaluate the contributions of the auxiliary OT and CL objectives in MASKED-UOT. As shown in Table 2, we compared four configurations: a baseline with BCE loss, OT only, CL only, and OT+CL, using MC3-18 (Tran et al., 2018) and Structured-ClinicalTokenEncoder. The transport-based fusion module remained active in all configurations;

Table 2: Loss-component ablation of MASKED-UOT on the real clinical dataset using MC3-18 and StructuredClinicalTokenEncoder. Values are mean ± sample standard deviation across four fold-trained models evaluated on the same held-out test partition. The transport-based fusion module remains active in every configuration; $\lambda_{\mathrm{OT}} = 0$ removes only the auxiliary OT loss. Bold indicates the highest mean in each metric column.

ID	Configuration	λ_{OT}	β_{prior}	λ_{CL}	AUC ↑	BACC ↑	F1 ↑
A	Baseline (BCE)	0	0	0	0.6631 ± 0.0267	0.5983 ± 0.0411	0.4870 ± 0.0585
B	BCE + OT-only	0.1	0.01	0	0.7041 ± 0.0585	$\mathbf{0.6384} \pm 0.0208$	0.5168 ± 0.0386
C	BCE + CL-only	0	0	0.1	0.6827 ± 0.0392	0.6325 ± 0.0720	$\mathbf{0.5433} \pm 0.0538$
D	BCE + OT + CL	0.1	0.01	0.1	$\mathbf{0.7176} \pm 0.0408$	0.6293 ± 0.0224	0.5197 ± 0.0434

setting $\lambda_{\mathrm{OT}} = 0$ therefore removed only the auxiliary OT loss. The OT-enabled configurations used $\lambda_{\mathrm{OT}} = 0.1$ and $\beta_{\mathrm{prior}} = 0.01$, while the CL-enabled configurations used $\lambda_{\mathrm{CL}} = 0.1$. Results are reported as mean ± sample standard deviation across four fold-trained models evaluated on the same held-out test partition.

The baseline achieved an AUC of 0.6631, BACC of 0.5983, and F1 of 0.4870. Adding the OT objective improved all three mean metrics, yielding an AUC of 0.7041, the highest BACC of 0.6384, and an F1 of 0.5168. Adding CL to BCE increases mean AUC to 0.6827 and BACC to 0.6325, while achieving the highest mean F1-score (0.5433). Combining OT and CL produced the highest AUC (0.7176 ± 0.0408), with BACC of 0.6293 ± 0.0224 and F1 of 0.5197 ± 0.0434.

These results highlight complementary performance trade-offs under the approximately 2:1 alive-to-deceased class imbalance. AUC assesses discrimination across decision thresholds, BACC weights sensitivity and specificity equally, and deceased-class F1 summarizes precision and recall without accounting for true negatives. Accordingly, the observed means favor OT+CL for risk ranking, OT alone for balanced classification, and CL alone for deceased-class F1.

4.5 PARAMETER SENSITIVITY ANALYSIS

We investigated the sensitivity of MASKED-UOT to three key hyperparameters: the OT-loss weight λ_{OT}, the prior-regularization weight β_{prior}, and the contrastive-loss weight λ_{CL}. In the BCE+OT-only configuration, increasing λ_{OT} and β_{prior} tenfold produced a modest AUC improvement of 0.35 percentage points, from 0.7041 to 0.7076. In the BCE+CL-only configuration, increasing λ_{CL} tenfold produced a larger improvement of 4.00 percentage points. In contrast, simultaneously increasing all three weights tenfold reduced the AUC from 0.7176 to 0.6749, a decrease of 4.27 percentage points. Furthermore, disabling prior regularization in the BCE+OT-only configuration reduced the AUC from 0.7041 to 0.6912, a decrease of 1.29 percentage points. Overall, these results indicate that performance is somewhat sensitive to the relative weighting of the auxiliary objectives and that increasing all loss weights does not necessarily improve performance.

5 CONCLUSION

We introduced MASKED-UOT, an availability-constrained UOT module for multimodal alignment and fusion. The method excludes unavailable observations through a hard support constraint, learns patient-specific priors over the remaining tokens, relaxes transported mass through KL penalties, and uses the resulting coupling for bidirectional feature transfer. In two lung-cancer cohorts, MASKED-UOT improves several ranking metrics over matched balanced-OT models, but the magnitude and direction of the gains depend on the encoder, cohort, and metric. This result is encouraging and also clarifies the next empirical step: controlled missingness, component-level ablations, stronger uncertainty analysis, and validation beyond CT-clinical prediction.

As part of future work, we plan to extend MASKED-UOT to handle entirely missing modalities, longitudinal observations, and time-to-event outcomes with censoring. We will also investigate component-level ablations, uncertainty-aware transport, and clinically validated correspondence maps, while evaluating the framework across larger multi-institutional cohorts and broader multimodal benchmarks to establish its robustness and generalizability.

AI USE STATEMENT

Large language models (LLMs), specifically OpenAI's ChatGPT and Anthropic's Claude, were used to assist in paper preparation and implement debugging. LLM outputs were carefully reviewed and edited for accuracy. Their role was limited to code debugging, language refinement, LaTeX formatting, and generating alternative phrasings of author-written content. All scientific ideas, experimental design, analysis, and claims were conceived, implemented, and verified by the authors. The authors take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the data preprocessing, patient-level partitioning, model architectures, training objectives, hyperparameter settings, and evaluation procedures used in our experiments in Section 4.1 and Appendix C. We report performance as the mean and sample standard deviation across four fold-trained models evaluated on the same held-out test partition, with classification thresholds selected using the corresponding validation partitions in Section 4.2. These standard deviations quantify variability across trained models rather than uncertainty from independently sampled test cohorts. All anonymous implementation code is included in the supplementary materials.

ETHICS STATEMENT

This retrospective study used de-identified clinical and imaging data under applicable institutional review board oversight. In addition, an open-source benchmark dataset (NSCLC-Radiomics (LUNG1) cohort (Aerts et al., 2014)) was used to evaluate the performance of the proposed method. All data were stored and analyzed in secure environments with access restricted to authorized personnel. MASKED-UOT is intended solely for research and does not provide autonomous clinical recommendations. Because retrospective cohorts may reflect demographic, institutional, and treatment-related biases, performance should be evaluated across clinically relevant subgroups and external populations before deployment. Transport weights indicate statistical associations rather than causal explanations, and predictions should not replace clinician judgment. Future clinical use will require prospective validation, privacy safeguards, uncertainty reporting, ongoing monitoring for distribution shift and inequitable performance, and appropriate human oversight.

REFERENCES

Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1):4006, 2014.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Jiequan Cui, Beier Zhu, Qingshan Xu, Zhuotao Tian, Xiaojuan Qi, Bei Yu, Hanwang Zhang, and Richang Hong. Generalized kullback-leibler divergence loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, and Marinka Zitnik. Multimodal learning with graphs. *Nature Machine Intelligence*, 5(4):340–350, 2023.

Meri Ferretti and Valentina DA Corino. Integrating radiomic and 3d autoencoder-based features for non-small cell lung cancer survival analysis. *Computer Methods and Programs in Biomedicine*, 258:108496, 2025.

Ronen Fluss, David Faraggi, and Benjamin Reiser. Estimation of the youden index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(4):458–472, 2005.

Yi Gu, Kuniaki Saito, and Jiaxin Ma. Learning contrastive multimodal fusion with improved modality dropout for disease detection and prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 280–290. Springer, 2025.

Wei Han, Hui Chen, Min-Yen Kan, and Soujanya Poria. Mm-align: Learning optimal transport-based alignment dynamics for fast and accurate inference on missing modality sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10498–10511, 2022.

Fangyu Hao, Jiayu Yang, Yifan Zhu, Zijun Yu, Qicen Wu, Wang Yunlong, Jiawei Li, Yulin Liu, Xu Zeng, Guanting Chen, et al. Lungcure: Benchmarking multimodal real-world clinical reasoning for precision lung cancer diagnosis and treatment. *arXiv preprint arXiv:2604.06925*, 2026.

Guolin Huang, Wenting Chen, Jiaqi Yang, Xinheng Lyu, Xiaoling Luo, Sen Yang, Xiaohan Xing, and Linlin Shen. Survagent: Hierarchical cot-enhanced case banking and dichotomy-based multi-agent system for multimodal survival prediction. *arXiv preprint arXiv:2511.16635*, 2025.

Elizabeth Huynh, Ahmed Hosny, Christian Guthier, Danielle S Bitterman, Steven F Petit, Daphne A Haas-Kogan, Benjamin Kann, Hugo JWL Aerts, and Raymond H Mak. Artificial intelligence in radiation oncology. *Nature Reviews Clinical Oncology*, 17(12):771–781, 2020.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s):1–41, 2022.

Kyungwon Kim, Yongmoon Lee, Doohyun Park, Taejoon Eo, Daemyung Youn, Hyesang Lee, and Dosik Hwang. Llm-guided multi-modal multiple instance learning for 5-year overall survival prediction of lung cancer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 239–249. Springer, 2024.

John Lee, Max Dabagia, Eva Dyer, and Christopher Rozell. Hierarchical optimal transport for multimodal distribution alignment. *Advances in Neural Information Processing Systems*, 32, 2019.

J Lekha, B Meena Preethi, KP Yuvaraj, and Angelo P Chery. Multimodal learning for autonomous systems and robotics. In *Multimodal Learning Using Heterogeneous Data*, pp. 193–215. Elsevier, 2026.

Tianyi Lin, Nhat Ho, and Michael I Jordan. On the efficiency of entropic regularized algorithms for optimal transport. *Journal of Machine Learning Research*, 23(137):1–42, 2022.

Vanessa Lopez, Lam Hoang, Marcos Martinez-Galindo, Raúl Fernández-Díaz, Marco Luca Sbodio, Rodrigo Ordonez-Hurtado, Mykhaylo Zayats, Natasha Mulligan, and Joao Bettencourt-Silva. Enhancing foundation models for scientific discovery via multimodal knowledge graph representations. *Journal of Web Semantics*, 84:100845, 2025.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

Zaiyu Pan, Shuangtian Jiang, Xiao Yang, Hai Yuan, and Jun Wang. Hierarchical cross-modal image generation for multimodal biometric recognition with missing modality. *IEEE Transactions on Information Forensics and Security*, 20:4308–4321, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):742–754, 2024.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

Kyle Swanson, Eric Wu, Angela Zhang, Ash A Alizadeh, and James Zou. From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell*, 186 (8):1772–1791, 2023.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459. IEEE, 2018.

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

Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1828–1838, 2020.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*, 112 (11):1716–1754, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Honglong Yang, Shanshan Song, Yi Qin, Lehan Wang, Haonan Wang, Xinpeng Ding, Qixiang Zhang, Bodong Du, and Xiaomeng Li. Multi-modal explainable medical ai assistant for trustworthy human-ai collaboration. *arXiv preprint arXiv:2505.06898*, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Haochen You and Baojing Liu. Mover: Multimodal optimal transport with volume-based embedding regularization. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5444–5448, 2025.

Yuan Yuan, Zhaojian Li, and Bin Zhao. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*, 57(7):1–34, 2025.

Chenyu Zhao, Yingxue Xu, Fengtao Zhou, Yihui Wang, and Hao Chen. Llm-driven knowledge enhancement for multimodal cancer survival prediction. *arXiv preprint arXiv:2512.14594*, 2025.

Yu Zhu, Liqiang Song, Guodong Wang, Junli Zhao, Hui Li, and Yi Li. Text-guided medical image fusion using unbalanced optimal transport: Semantic alignment and cross-modal interaction. *Signal Processing*, pp. 110590, 2026.