AlignedFusion: Handling Missing Information in Reports and Inter-Modality Information Imbalance

Han Li[1,2,3] TUM_HAN.LI@TUM.DE
Jingsong Liu[2,3] JINGSONG.LIU@TUM.DE
Zhengyang Xu[2,3] ZHENGYANG.XU@TUM.DE
Peter Schüffler[2,3] PETER.SCHUEFFLER@TUM.DE
Nassir Navab[1,2] NASSIR.NAVAB@TUM.DE
S.Kevin Zhou[*4] SKEVINZHOU@USTC.EDU.CN

[1] *Computer Aided Medical Procedures (CAMP), Technische Universitaet Muenchen (TUM).*
[2] *Munich Center for Machine Learning (MCML), Munich, Germany.*
[3] *Institute of Pathology, TUM School of Medicine and Health, Technical University of Munich (TUM), Germany*
[4] *Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE), Suzhou Institute for Advance Research, USTC, Suzhou, 215123, China*

Editors: Under Review for MIDL 2026

Abstract

Integrating textual reports and visual information is crucial for multimodal medical AI. However, existing approaches face two major challenges: (1) Handling omitted information in reports, as textual encoders struggle to differentiate between unmentioned and truly absent attributes, leading to inconsistencies in feature learning, and (2) Inter-modality information imbalance, where direct token-wise attention between text and images causes instability due to the disparity in information richness between modalities. To address these issues, we propose **AlignedFusion**, a novel multimodal fusion framework with two key components: (1) Attribute-wise Report Token Generation with Masked Token Reconstruction, which structures medical reports into explicit attribute categories and reconstructs missing attributes to reduce feature variance, and (2) Intermediate Token-Based Fusion, which stabilizes multimodal learning by inserting an intermediate token as a bridge between textual and visual representations, ensuring a balanced and effective fusion. We evaluate AlignedFusion on four medical analysis tasks using two public and two private datasets, demonstrating its adaptability and robustness. Experimental results show that our approach improves alignment between textual and visual features, mitigates training instability, and enhances predictive performance, advancing the field of multimodal medical AI. Code will be available upon acceptance.

Keywords: Missing data, Transformer, Pathology, CT lesion detection, Skin tumor

1. Introduction

Clinical reports (Downing, 2001) contain rich diagnostic information that extends beyond what is visible in medical images (deSouza et al., 2019; Bayer, 2018; O'connor et al., 2017; Hait, 2011; La Thangue and Kerr, 2011). Numerous medical diagnostic algorithms have

* Corresponding Author

demonstrated the critical role of these textual reports in machine learning applications, showing their effectiveness in improving predictive accuracy and clinical decision-making (Mohsen et al., 2022; Tschandl, 2020; Tang, 2022; Braman et al., 2021; Zuo et al., 2022; Cui et al., 2022; Guan et al., 2021; Chauhan et al., 2020; Silva and Rohr, 2020; Qiang et al., 2021). As a result, integrating textual report and visual information has become a crucial research direction in multimodal learning for medical AI. Recently, the most popular strategy is to first use a textual encoder (e.g., BERT (Devlin et al., 2019), ClinicalBERT (Huang et al., 2019), or BioBERT (Lee et al., 2020)) and a visual encoder (e.g., transformer-based (Dosovitskiy et al., 2020) or convolutional methods) to encode the report and medical image, respectively, and then integrate the features for the final task. As shown in Fig.1, while achieving success, these methods still face two major challenges: **1) Inability to Handle Omitted Information in Reports.** Reports vary in detail, some attributes are explicitly recorded, while others may be omitted, making it unclear whether an attribute is truly absent or simply deemed unnecessary to mention. Textual encoders cannot effectively handle missing attributes because they do not know which attributes are omitted, leading to inconsistencies during training, where certain attributes appear in some reports but are absent in others. This increases feature variance in report representations, making it more challenging to align textual and visual features. Additionally, some missing attributes could be easily inferred from other available attributes, yet current models fail to leverage this contextual knowledge. **2) Training Instability Due to Inter-Modality Information Imbalance.** Medical reports primarily describe the presence or absence of specific attributes, containing significantly less information than images. However, current methods train textual and visual features together using direct token-wise attention, despite the substantial difference in the amount of information each modality provides. This imbalance can make joint training unstable, causing the model to either overfit to the more information-rich visual features or fail to effectively utilize the comparatively limited textual information.

To address these issues, we propose AlignedFusion to handle missing information in reports and inter-Modality information imbalance. It consists of two key components: **1) Attribute-wise Report Token Generator and Masked Token Reconstruction.** We use Attribute-wise Report Token Generator to structure the medical report into explicit attribute categories, classifying each attribute as positive, negative, or unmentioned. The selected attributes can either be manually predefined or extracted from the report cohort of the entire dataset. To handle the unmentioned attributes, we introduce a Masked Token Reconstruction model, treating them as unknown masks and training the model to reconstruct them based on available data. This prevents the model from incorrectly assuming unmentioned attributes are negative while enabling a data-driven inference of missing information. **2) Intermediate Token-Based Fusion.** To mitigate the instability caused by direct token-wise attention between text and images, we introduce an Intermediate Token-based fusion mechanism as a narrow bridge between modalities. Instead of directly applying attention across all tokens, we insert an intermediate token between each modality pair, training it to encode only the shared information between them. This prevents high-density image features from overwhelming sparse textual attributes, ensuring that only the most critical cross-modal information is preserved.

To better demonstrate the efficacy of our AlignedFusion, we conduct extensive experiments on four representative medical analysis tasks spanning both public and private

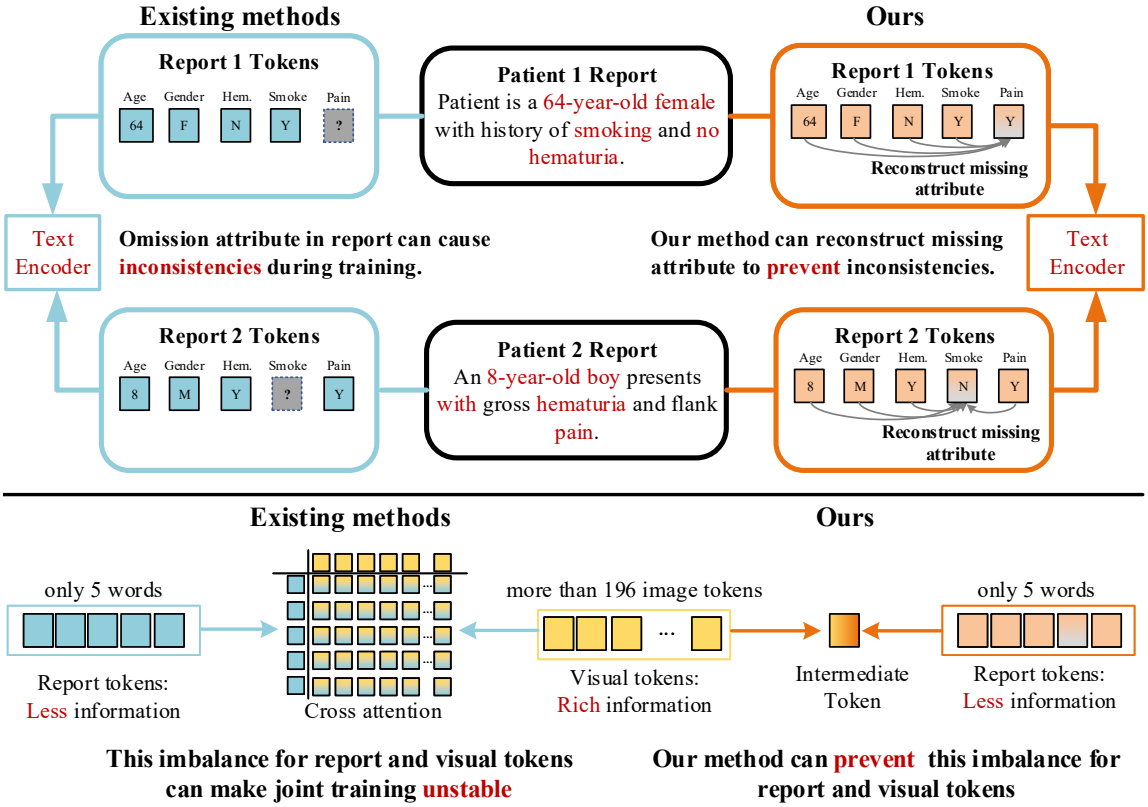

Figure 1: UPPER: Current methods do not know which attributes are omitted, leading to inconsistencies during training, where certain attributes appear in some reports but are absent in others. In contrast, our method can prevent the model from incorrectly assuming unmentioned attributes. LOWER: Current methods train less textual and rich visual features together using direct token-wise attention, despite the substantial difference in the amount of information each modality provides. In contrast, our method introduces an intermediate token to mitigate this imbalance.

datasets. Specifically, we evaluate our method on two widely used public and two in-house clinical datasets, covering diverse imaging modalities(CT, pathology, and dermoscopic images), disease types, and reporting styles. Across all tasks, we systematically simulate different levels of report incompleteness. The results consistently show that AlignedFusion achieves superior performance and exhibits clear robustness under varying missing-attribute ratios, highlighting its practical applicability to real-world clinical scenarios where textual reports are often sparse, incomplete, or heterogeneous.

2. Method

As shown in Fig.2, we hereby take the two-modality-input (one image modality and one report modality) as a working example to illustrate how we tackle missing attributes, but it supports more modality inputs. We will first introduce four successive parts of our methods

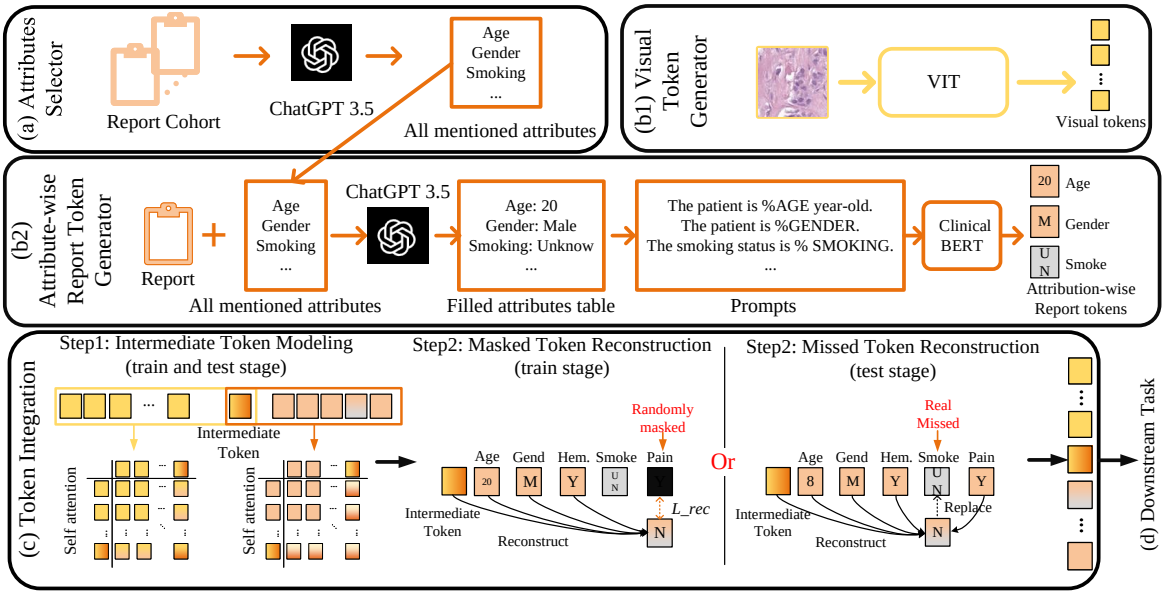

Figure 2: The network architecture.(a) Attribute Selector: Automatically extracts a unified set of clinically relevant attributes from the report cohort using GPT-based prompting. (b1) Visual Token Generator: Encodes the input image into patch-level visual tokens with a pretrained Vision Transformer. (b2) Attribute-wise Report Token Generator: Represents each attribute as a dedicated report token and assigns a special UN token to attributes not mentioned in the report. (c) Token Integration: Integrates visual and report tokens via an intermediate-token based fusion with masked token reconstruction to mitigate missing attributes. (d) Downstream Task: Uses the fused token representation for task-specific prediction, such as image-level classification.

in 2.1 to 2.4 and then describe how to downstream out model in 2.5. Finally, we present the training and testing procedures in 2.6.

2.1. Attribute Selector F_{AS}

For a dataset $\boldsymbol{X}$ containing n image-report pairs (i.e., two modalities):

$$\boldsymbol{X} = [\boldsymbol{I}, \boldsymbol{R}], \quad \boldsymbol{I} = [I^1, ..., I^n], \quad \boldsymbol{R} = [R^1, ..., R^n] \tag{1}$$

we first extract all mentioned attributes names N_A from the reports using ChatGPT-3.5:

$$N_A = F_{AS}(\boldsymbol{R}) = \text{Chat}(\boldsymbol{R}, \text{Prompt}), \quad N_A = [N_{a^1}, ..., N_{a^i}] \tag{2}$$

where i is the total number of selected attributes, and Prompt is a query in the format: *"Please list all mentioned attributes which are important for %TASK."* Here, %TASK refers to the specific downstream task. We removed the superscript of images and reports in the following for simplification.

2.2. Token Generators G

This part includes two sub parts: Visual Token Generator G_{vis} and Attribute-wise Report Token GeneratorG_{rep}. The visual tokens and Report tokens are generated by Visual Token Generator and Attribute-wise Report Token Generator respectively.

Visual Token Generator G_{vis}. We adopt a pretrained ViT(Liu et al., 2021) F_{ViT} as the backbone to extract the image visual tokens $t_v = [t_v^1, ..., t_v^i]$ from images I:

$$[t_v^1, ..., t_v^i] = F_{ViT}(P), \quad P = [P^1, ..., P^i] = F_P(I), \tag{3}$$

where the P represents the image patches generated from the partial operations F_P, and i is the number of patches.

Attribute-wise Report Token Generator G_{rep}. We first utilize GPT-3.5 $Chat(\cdot)$to extract the values of different attributes N_A from the report R. If an attribute is explicitly mentioned in R, we record its actual value; otherwise, we define it as **UN** (Unknown):

$$Chat(R, N_{a^i}) = \begin{cases} a^i, & \text{if } N_{a^i}\text{is mentioned in } R \\ \text{UN}, & \text{if } N_{a^i} \text{ is not mentioned in } R \end{cases} \tag{4}$$

where a^i represents the extracted value of attribute N_{a^i} from the report. With this equation we can get a filled attributes table for report R: $Tab_A = \{N_{a^i} : Chat(R, N_{a^i})\}$.

After obtaining the filled attributes table Tab_A^n, we construct prompts for each attribute $Pt_{N_{a^i}}$, and pass them separately through ClinicalBERT (Huang et al., 2019) G_{cBERT} to obtain attribute-level tokens. If the attribute value is **Unknown (UN)**, we directly replace it with a predefined **UN mask token** t_{UN} without processing it through ClinicalBERT:

$$t_r^i = \begin{cases} G_{cBERT}(Pt_{N_{a^i}}), & \text{if } a^i \neq \text{UN} \\ t_{UN}, & \text{if } a^i = \text{UN} \end{cases} \tag{5}$$

where t_n^i represents the attribute-level token embedding, extracted from the [CLS] token of ClinicalBERT. As shown in Fig.2 (b2), our prompt $Pt_{N_{a^i}}$ design is straightforward, and our experiments indicate that the choice of prompt design has a negligible impact on performance. Finally, we aggregate all attribute-level tokens to form the Attribute-wise Report Token: $t_r = \{t_r^1, ..., t_r^i\}$.

2.3. Token Integration

As shown in Fig.2 (c), after obtaining the Visual and Attribute-wise Report Tokens, we integrate them for the downstream task. As discussed in the Introduction, we first introduce an Intermediate Token to mitigate the instability caused by direct token-wise attention between text and images. Then, we apply Masked Token Reconstruction to handle unmentioned tokens.

2.4. Intermediate Token t_{int} Modeling

To serve as a bridge, the same Intermediate Token t_{int} is added to both modality token sets:

$$t'_v = [t_v; t_{(v,int)}], \quad t'_r = [t_{(r,int)}; t_r], \quad t_{int} = t_{(v,int)} = t_{(r,int)} \tag{6}$$

For each token set, all tokens interact through the self-attention operation F_{SA}:

$$t'^{SA}_{v} = F_{SA}(t'_v), \quad t'^{SA}_{r} = F_{SA}(t'_r). \tag{7}$$

Since the Intermediate Token is modeled twice, once in each modality, we integrate its two representations using an averaging operation Avg:

$$t^{SA}_{int} = \text{Avg}(t^{SA}_{(v,int)}, t^{SA}_{(r,int)}), \quad t^{SA}_{int} \rightarrow t^{SA}_{(v,int)}, \quad t^{SA}_{int} \rightarrow t^{SA}_{(r,int)}. \tag{8}$$

Now, the Intermediate Token captures information from both modalities. In other words, even when using the report token set t'^{SA}_{r}, it can still access visual information through the Intermediate Token.

Masked Token Reconstruction model F_{rec}. Within the report token set, the Masked Token Reconstruction model F_{rec} first randomly masks out tokens with the UN mask at a certain probability $\boldsymbol{p}$, mimicking unmentioned attributes. It then learns to reconstruct the masked tokens based on the remaining tokens within the report token set:

$$t^i_r = \begin{cases} t^i_r, & \text{if } a \geq p \\ t_{UN}, & \text{if } a < p \end{cases} \quad a = Rand(0,1). \tag{9}$$

We empirically set p=0.3 in our experiments. Then the reconstruct loss is calculated based on the masked tokens only:

$$\hat{t}^i_r = F_{rec}(t'^{SA}_{r} - t^i_r) = F_{rec}, l_{rec} \quad = \text{CE}(\hat{t}^i_r, t^i_r). \tag{10}$$

2.5. Downstream model F_d

Taking disease classification as an example, the token sets t'^{SA}_{r} and t'^{SA}_{v} are concatenated to generate a global representation t_{cls}. This global representation t_{cls} is then used by the downstream model F_d to predict the final classification result $\hat{Y}$, optimized with the loss function l_d:

$$\hat{Y} = F_d(t_{cls}), \quad l_d = L_{CE}(Y, \hat{Y}). \tag{11}$$

2.6. Overall training and testing

During training, the total loss is the sum of the downstream task loss and the masked token reconstruction loss: $l_{all} = l_d + l_{rec}$.

For unmentioned attributes during training, we directly use the reconstruction network to reconstruct their tokens for subsequent training.

During testing, as shown in Fig.2 (c), the Masked Token Reconstruction is replaced by Missed Token Reconstruction. No masking is applied; instead, the model reconstructs only the genuinely unmentioned attributes.

3. Experiments

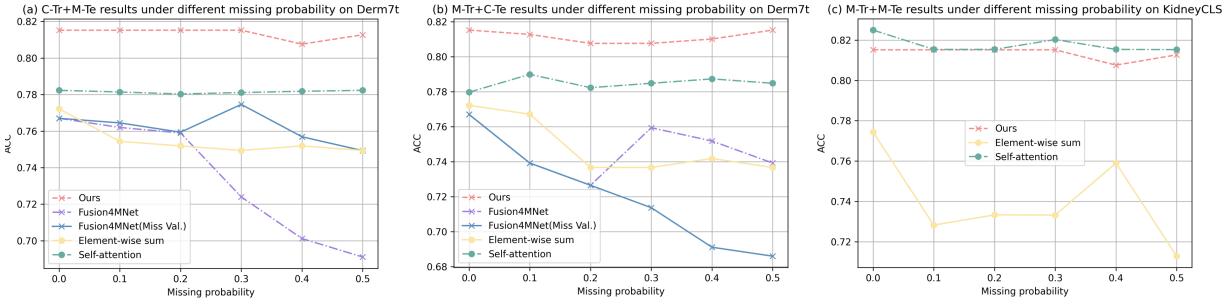

Figure 3: The line graph of classification accuracy (%) under the Attribute M-Tr+C-Te and C-Tr+M-Te scenarios on the Derm7t(Kawahara et al., 2019) and KidneyCLS datasets.

3.1. Dataset and setting

Our experiments are conducted on four datasets, including two public and two private datasets. The public dataset Derm7pt (Kawahara et al., 2019) contains 413 training cases, 203 validation cases, and 395 testing cases. Each case comprises a dermoscopic image and a clinical image, with five selected attributes: Mel, Nevus (Nev), Seborrheic Keratosis (SK), BCC, and Miscellaneous (Misc). The public dataset Deeplesion (Yan et al, 2018) is designed for CT lesion detection and contains 32,120 axial CT slices. Four attributes are extracted from the report: patient age, lesion location, gender, and lesion size. We compare our method with the previous leading method, FusionM4Net(Tang, 2022). The private dataset, KidneyCLS, was collected from an in-house hospital between 2012 and 2020 and was annotated by three doctors with at least five years of clinical experience. The dataset contains 648 Kidney CT images, with each case including a CT image and 13 attributes extracted from the report. The private dataset, Skin Tumor Dataset designed for patch-level Skin Tumor classification, consists of 120 skin tumor pathology slides (114k annotated patches), with 14 attributes extracted from the report.

3.2. Performance in different attribute missing scenarios

First, to evaluate performance under different attribute missing scenarios, we conduct experiments in Attribute"M-Tr+C-Te" (Missing Training and Complete Testing), "C-Tr+M-Te"(Complete Training and Missing Testing), and "M-Tr+M-Te" (Missing Training and Missing Testing) scenarios and report results in Fig. 3 and Table 1.

Attribute M-Tr+C-Te results. We learn separate models with different attribute missing rates in the training set and evaluate them using attribut-complete test data. As in Table 1(a) and Fig. 3(a), our method has the best biomarker-complete test performances for all conditions. For example, with a 20% attribute missing rate in training, our model achieves an accuracy of 80.76%, exceeding the competing method FusionM4net by a margin of 8.11%.

Attribute C-Tr+M-Te results. As in Fig. 3(b) and Table 1(b), our method learned with an attribute-complete training set exhibits accuracy resilience in the face of progressively incomplete test datasets; After a slight decline, performance interestingly shows a tendency to recover. This phenomenon can be attributed to the strategic of our method to data

Table 1: Classification accuracy (%) under the AttributeM-Tr+C-Te, C-Tr+M-Te, and M-Tr+M-Te scenarios on the Derm7t (Kawahara et al., 2019) and KidneyCLS datasets.

	(a) M-Tr+C-Te — Derm7t dataset					
Methods	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**
Ours	**81.52**	81.27	80.76	80.76	81.01	81.52
FusionM4net(Tang, 2022)	76.70	73.92	72.65	75.94	75.18	73.92
FusionM4net(Tang, 2022)(Miss Val.)	76.70	73.92	72.65	71.37	69.11	68.60
Element-wise Sum	77.22	76.71	73.67	73.67	74.18	73.67
Self-Attention	77.97	78.99	78.23	78.48	78.73	78.48
RemixFormer(Xu et al., 2022)	81.30	-	-	-	-	-
MSMA(Shu et al., 2024)	78.99	-	-	-	-	-
Ours (Img Only)	76.55	-	-	-	-	-
FusionM4net(Tang, 2022) (Img Only)	75.40	-	-	-	-	-
	(b) C-Tr+M-Te — Derm7t dataset					
Methods	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**
Ours	81.52	81.27	80.76	80.76	81.01	81.52
FusionM4net(Tang, 2022)	76.70	76.45	75.94	77.46	75.69	74.93
FusionM4net(Tang, 2022)(Miss Val.)	76.70	76.20	75.90	72.40	70.12	69.11
Element-wise Sum	77.22	75.44	75.19	74.94	75.19	74.94
Self-Attention	78.23	78.14	78.03	78.11	78.18	78.23
RemixFormer(Xu et al., 2022)	81.30	-	-	-	-	-
MSMA(Shu et al., 2024)	78.99	-	-	-	-	-
Ours (Img Only)	76.55	-	-	-	-	-
FusionM4net(Tang, 2022) (Img Only)	75.40	-	-	-	-	-
	(c) M-Tr+M-Te — Derm7t dataset					
Methods	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**
Ours	**81.52**	**81.52**	**81.27**	**81.01**	**81.52**	**81.77**
Element-wise sum	-	-	-	-	-	-
Self-Attention	78.48	78.48	78.48	78.73	78.73	78.73
	(d) M-Tr+M-Te — KidneyCLS dataset					
Methods	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**
Ours	**82.73**	81.03	**82.56**	81.54	**82.56**	80.51
Element-wise sum	77.44	72.82	73.33	73.32	75.90	71.28
Self-Attention	82.05	**81.54**	81.54	**82.03**	81.54	**81.53**

incompleteness, wherein it engages in the reconstruction of attributes' soft labels, infused with enhanced information.

Attribute M-Tr+M-Te results. In this study, we conduct an experiment where the training set is subjected to a 50% missing rate of attributes, and the model's performance is subsequently evaluated on test sets experiencing various degrees of attribute loss on both datasets. The results are in Table 1(c,d) and Fig. 3(c). The outcomes of this experiment reveal that our methodology is capable of undergoing both training and testing phases effectively, even when a significant portion of the dataset is missing.

3.3. Effectiveness of Attribute-wise Report Token Generator

As shown in Table 2, our method achieves the best performance on the Skin tumor dataset for all missing-attribute ratios. When no attribute is missing in the training reports (0%),

Table 2: Classification accuracy (%) under the Attribute M-Tr+C-Te scenario on the Skin tumor dataset dataset.

	(a) M-Tr+C-Te — Skin tumor dataset					
Methods	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**
Element-wise Sum	70.22	69.71	68.67	67.67	66.18	65.67
Self-Attention	71.70	69.92	68.65	67.94	67.18	66.92
FusionM4net(Tang, 2022)	77.97	78.99	78.23	78.48	78.73	78.48
Ours +ClnicalBERT(Huang et al., 2019)	74.27	71.26	69.14	68.57	68.07	65.57
Ours +BioBERT(Lee et al., 2020)	76.24	73.14	71.56	70.65	69.99	66.35
RemixFormer(Xu et al., 2022)	79.30	-	-	-	-	-
MSMA(Shu et al., 2024)	77.69	-	-	-	-	-
Ours	81.52	81.27	80.76	80.76	81.01	81.52

Table 3: Attribute C-Tr+C-Te scenario: Sensitivity (%) of AlignedFusion at various FPPI on the official testing dataset of DeepLesion (Yan et al, 2018) under 25%, 50 % and 100 % training data.

Method	**data**	**slice**	@0.5	@1	@2	Avg.[0.5,1,2]
A3D(Yang et al, 2021)	25%	7	55.67	65.39	73.35	64.80
A3D+Ours	25%	7	57.04 (+1.37)	66.55 (+1.16)	73.97 (+0.62)	65.85 (+1.05)
A3D(Yang et al, 2021)	50%	7	72.52	80.27	86.14	79.64
A3D+Ours	50%	7	73.48 (+0.96)	80.87 (+0.60)	86.43 (+0.29)	80.26 (+0.62)
A3D(Yang et al, 2021)	100%	7	79.24	85.04	89.15	84.48
A3D+Ours	100%	7	80.21 (+0.97)	86.04 (+1.00)	89.66 (+0.51)	85.30 (+0.82)

simple fusion baselines such as Element-wise Sum and Self-Attention obtain 70.22% and 71.70% accuracy, respectively, while stronger multimodal baselines FusionM4net, MSMA, and RemixFormer reach 77.97%, 77.69%, and 79.30%. In comparison, our method attains 81.52% accuracy in the same setting.

As the missing-attribute ratio increases from 0% to 50%, the performance of Element-wise Sum and Self-Attention gradually decreases to 65.67% and 66.92%. Similarly, replacing our generator with ClinicalBERT or BioBERT leads to a clear degradation: the accuracies drop from 74.27% and 76.24% at 0% missing to 65.57% and 66.35% at 50% missing, indicating that they cannot effectively handle incomplete reports. In contrast, our method remains stable around 80–81% across all missing-attribute settings (from 81.52% at 0% to 81.52% at 50%), demonstrating that the proposed Attribute-wise Report Token Generator can effectively cope with incomplete reports and provides robust representations even when a large proportion of attributes are missing.

Table 4: Ablation study for (a) IT and (b) mask probabilities p in masked reconstruction model.

(a) IT	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**
Ours w/o IT	77.97	78.99	78.23	78.48	78.73	78.48
Ours	81.52	81.27	80.76	80.76	81.01	81.52
(b)p	$p = 0$	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$
Ours	81.01	81.04	80.25	81.52	80.07	0.7975

3.4. Effectiveness of Intermediate Token-Based Fusion

As shown in Table 3, when using our Intermediate Token-Based Fusion, it consistently improves sensitivity on DeepLesion across all data regimes and FPPI operating points. With only 25% of the training data, the average sensitivity over $\{0.5, 1, 2\}$ FPPI increases from 64.80% to 65.85% (+1.05), and similar gains are observed at each FPPI level. When using 50% of the data, our module boosts the average sensitivity from 79.64% to 80.26% (+0.62), and with 100% of the data, the average sensitivity further improves from 84.48% to 85.30% (+0.82). These consistent margins confirm that the proposed fusion strategy is an effective plug-and-play component that enhances lesion detection performance irrespective of the available training data size.

3.4.1. Ablation study

An ablation study is conducted to evaluate the importance of the two key components in our framework: (i) the Intermediate Token (IT) and (ii) the masked reconstruction model. As shown in Table 4(a), removing IT from SMF consistently leads to performance drops in the M-Tr+C-Te experiment across all missing-attribute settings. For example, at 0% and 50% missing rates, the accuracy decreases from 81.52% to 77.97% and from 81.52% to 78.48%, respectively. Similar gaps of around 2–3 percentage points are observed for the other missing probabilities, indicating that the intermediate token provides a more stable and balanced cross-modal interaction.

Table 4(b) further investigates the effect of the mask probability p in the masked reconstruction model. The performance is relatively stable for small to moderate masking rates, but a clear peak is achieved at $p = 0.3$, where the accuracy reaches 81.52%. When p is either too small ($p = 0$) or too large ($p = 0.5$), the accuracy slightly decreases.

4. Conclusion

We propose **AlignedFusion**, a novel framework that enhances multimodal fusion by introducing (1) Attribute-wise Report Token Generation with Masked Token Reconstruction to structure report attributes and infer missing information, and (2) Intermediate Token-Based Fusion to stabilize cross-modal attention and balance textual and visual contributions. We evaluate AlignedFusion on four medical analysis tasks across two public and two private datasets, demonstrating its adaptability and robustness. Our method improves alignment between textual and visual features while mitigating training instability, advancing multimodal learning for medical AI.

References

Antony J Bayer. The role of biomarkers and imaging in the clinical diagnosis of dementia. *Age and ageing*, 47(5):641–643, 2018.

Nathaniel Braman, Jacob WH Gordon, Emery T Goossens, Caleb Willis, Martin C Stumpe, and Jagadish Venkataraman. Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In *MICCAI*, pages 667–677. Springer, 2021.

Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *MICCAI*, pages 529–539. Springer, 2020.

Can Cui, Han Liu, Quan Liu, Ruining Deng, Zuhayr Asad, Yaohong Wang, Shilin Zhao, Haichun Yang, Bennett A Landman, and Yuankai Huo. Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. In *MICCAI*, pages 626–635. Springer, 2022.

Nandita M deSouza, Eric Achten, Angel Alberich-Bayarri, Fabian Bamberg, Ronald Boellaard, Olivier Clément, Laure Fournier, Ferdia Gallagher, Xavier Golay, Claus Peter Heussel, et al. Validated imaging biomarkers as decision-making tools in clinical trials and routine practice: current status and recommendations from the eiball* subcommittee of the european society of radiology (esr). *Insights into imaging*, 10:1–16, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

G Downing. Biomarkers definitions working group. biomarkers and surrogate endpoints. *Clin Pharmacol Ther*, 69:89–95, 2001.

Yulu Guan, Hui Cui, Yiyue Xu, Qiangguo Jin, Tian Feng, Huawei Tu, Ping Xuan, Wanlong Li, Linlin Wang, and Been-Lirn Duh. Predicting esophageal fistula risks using a multimodal self-attention network. In *MICCAI*, pages 721–730. Springer, 2021.

William N Hait. Forty years of translational cancer research. *Cancer Discovery*, 1(5): 383–390, 2011.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE J. Biomed. Health Inform.*, 23, 2019. doi: 10.1109/JBHI.2018.2824327. URL https://doi.org/10.1109/JBHI.2018.2824327.

Nicholas B La Thangue and David J Kerr. Predictive biomarkers: a paradigm shift towards personalized cancer medicine. *Nature reviews Clinical oncology*, 8(10):587–596, 2011.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

Farida Mohsen, Hazrat Ali, Nady El Hajj, and Zubair Shah. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Scientific Reports*, 12 (1):17981, 2022.

James PB O'connor, Eric O Aboagye, Judith E Adams, Hugo JWL Aerts, Sally F Barrington, Ambros J Beer, Ronald Boellaard, Sarah E Bohndiek, Michael Brady, Gina Brown, et al. Imaging biomarker roadmap for cancer studies. *Nature reviews Clinical oncology*, 14(3):169–186, 2017.

Mengyun Qiang, Chaofeng Li, Yuyao Sun, Ying Sun, Liangru Ke, Chuanmiao Xie, Tao Zhang, Yujian Zou, Wenze Qiu, Mingyong Gao, et al. A prognostic predictive system based on deep learning for locoregionally advanced nasopharyngeal carcinoma. *JNCI*, 113 (5):606–615, 2021.

Ci Shu, Long Yu, Shengwei Tian, and Xianwei Shi. Msma: A multi-stage and multi-attention algorithm for the classification of multimodal skin lesions. *BSPC*, 93:106180, 2024.

Luís A Vale Silva and Karl Rohr. Pan-cancer prognosis prediction using multimodal deep learning. In *ISBI*, pages 568–571. IEEE, 2020.

P. Tang. Fusionm4net: a multi-stage multi-modal learning algorithm for multi-label skin lesion classification. *MIA.*, 76, 2022.

P. Tschandl. Human-computer collaboration for skin cancer recognition. *Nat. Med.*, 26, 2020. doi: 10.1038/s41591-020-0942-0. URL https://doi.org/10.1038/s41591-020-0942-0.

Jing Xu, Yuan Gao, Wei Liu, Kai Huang, Shuang Zhao, Le Lu, Xiaosong Wang, Xian-Sheng Hua, Yu Wang, and Xiang Chen. Remixformer: A transformer model for precision skin tumor differential diagnosis via multi-modal imaging and non-imaging data. In *MICCAI*, pages 624–633. Springer, 2022.

K. Yan et al. Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In *IEEE/CVF CVPR*, pages 9261–9270, 2018.

J. Yang et al. Asymmetric 3d context fusion for universal lesion detection. In *MICCAI*, pages 571–580. Springer, 2021.

Yingli Zuo, Yawen Wu, Zixiao Lu, Qi Zhu, Kun Huang, Daoqiang Zhang, and Wei Shao. Identify consistent imaging genomic biomarkers for characterizing the survival-associated interactions between tumor-infiltrating lymphocytes and tumors. In *MICCAI*, pages 222–231. Springer, 2022.