Double Vision: Unifying Morphology and Gene Expression with a Multimodal Transformer

Anonymous Author(s)
Affiliation
Address
email

Abstract

Tissues can be characterized by their complex morphological structures and molecular programs, as captured by histology images and spatial transcriptomic technologies. Current unimodal foundation models are limited in their ability to reason across morphological and molecular features. We introduce a multimodal transformer architecture that unifies histology images and spatial transcriptomics through token-level fusion. By representing both modalities as interoperable tokens within a shared sequence, our model integrates morphological and molecular features throughout all layers, prioritizing cross-modal relationships over isolated single-modality representations. The resulting token-fusion transformer captures rich morphological and molecular signatures, contextualizing histopathology patterns with molecular information and vice versa. Though preliminary, our results demonstrate that token fusion enhances disease-state prediction and lay the groundwork for multimodal models capable of reasoning jointly over tissue morphology and gene expression.

1 Introduction

Tissue organization arises from the coordinated arrangement of many different cell types, each with distinct morphologies, phenotypes, and molecular programs [1]. Histopathology has long relied on hematoxylin and eosin (H&E) staining of whole-slide images (WSIs), which captures rich morphological information across tissue scales, from single cells to global organization. The emergence of Vision Transformers (ViTs) [2] as powerful image encoders has revolutionized computational pathology: ViTs operate on image patches as input tokens, capturing contextual dependencies across tissue scales. Self-supervised pretraining of ViTs on hundreds of millions of image patches from millions of H&E WSIs has led to general-purpose foundation models (FMs) such as UNI [3], Virchow [4] and Midnight [5], that excel across diverse pathology tasks, underscoring the power of morphological representations. In parallel, FMs for single-cell transcriptomics (sc-FMs, e.g., GeneFormer [6], scGPT [7]) provide rich molecular information across tissues and diseases. However, scFMs lack spatial resolution and therefore cannot directly connect molecular programs back to the tissue. Emerging spatial transcriptomics (ST) technologies promise to fill this gap: they enable the deep molecular profiling of individual cells within intact tissue [8]. Imaging-based ST such as Xenium (10x Genomics) exemplify this advance: they map millions of transcripts *in situ* at subcellular resolution, potentially exposing tissue niches and cellular interactions in health or disease [9].

Together, these developments have laid the groundwork for models that learn unified representations combining morphology and gene expression, with the potential of capturing a holistic view of tissue ecosystems. Recent efforts in this direction mostly follow the paradigm of CLIP (Contrastive Language-Image Pretraining) [10], and typically train dual encoders to align image with omics features in a shared latent space. For instance, TANGLE [11] learns slide-level embeddings by

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

contrasting H&E images with their bulk transcriptomic profile. In OmiCLIP [12], patches from H&E WSIs are paired with corresponding gene expression "sentences", and separate encoders–a ViT for images and a text encoder for the sentences–are trained with a contrastive objective. However, in all CLIP-like models, the modalities are *not* fused at any point during training. Instead, the transcriptomic data serve mainly as a supervisory signal to improve the image encoder via contrastive alignment. As such, these models remain fundamentally unimodal in their architecture and the representations from the different modalities are only merged late at inference time if at all, limiting their utility for tasks requiring unified morphological and molecular reasoning.

Here we introduce a new multimodal model that overcomes these limitations by performing *token-level fusion* of H&E images and ST features within a unified transformer architecture. In our approach, the two data modalities are merged *early*, as interoperable tokens in the same sequence, enabling the model to attend to joint morpho–molecular patterns at every layer. By building on powerful pretrained encoders per modality, we leverage prior learning – the image branch starts from a ViT model already trained on H&E images, and the ST branch employs a sc-FM – and focus on learning cross-modal relationships. The result is a flexible, token-fusion transformer that can enrich histological patterns with gene expression context and vice versa, capturing unified signatures.

2 Methods

Model Architecture Our model (Figure 1) builds upon any pretrained ViT on H&E images (e.g., UNIv2 [3], Midnight [5]) as a unified encoder for both H&E image and ST data. Both H&E images and ST data are "tokenized" in a consistent way based on their coordinates to create a set of spatially aligned image / transcript tokens, which, together with the corresponding positional information for each token, is encoded by the pretrained ViT. This design allows seamless switching between unimodal and multimodal inference: the model can ingest an image alone (using image tokens only), an ST sample alone, or both together in an integrated fashion.

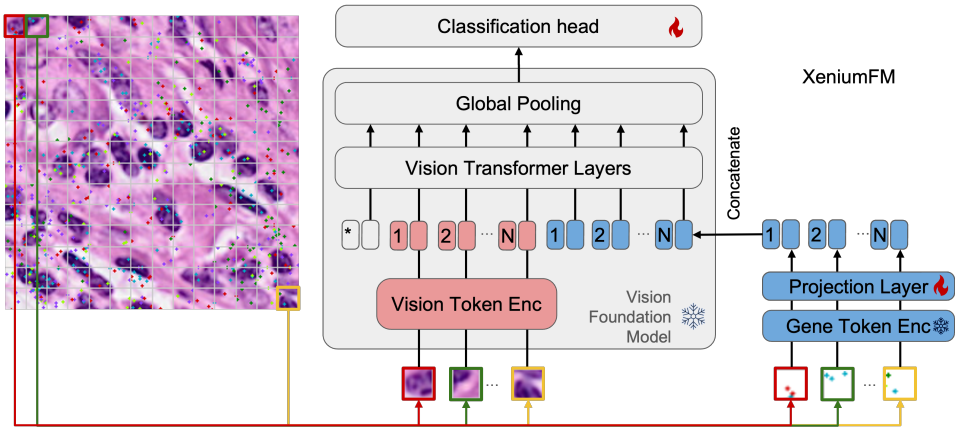

Figure 1: A novel multimodal transformer architecture fusing H&E and ST images: The H&E and ST images are spatially aligned and tokenized. H&E tokens are embedded using the patch embedding layer in the vision-only pathology FM, while transcript tokens are embedded using a sc-FM. The resulting tokens are concatenated before being passed through a transformer encoder. The vision encoder and the gene encoder are kept frozen during training (snowflake), while the projection layer and classification head are trained for each downstream task (flame).

Tokenization The tokenization of images follows the standard ViT patching scheme as is usually defined in the patch embedding layer. Input images are divided into fixed-size patches: e.g. a 224×224 pixel region is split into a 14×14 grid of patches with 16×16 pixels each, yielding 196 image tokens. We also keep the usual *cls* token from the ViT architecture, which aggregates the fused information for downstream predictions. For the ST modality, we develop an analogous tokenization strategy to represent spatial gene expression in a ViT-compatible manner. In the Xenium images, each detected transcript is associated with gene id G_i and tissue coordinates (x_i, y_i). To align those to the H&E image, we partition the ST image into patches corresponding to the H&E image patch grid–the

same 14×14 layout over the tissue area. For each region R_j, we aggregate the set of local transcripts $(x_i, y_i, G_i) \mid i \in R_j$ into a vector of gene expression vector $\mathbf{M}_j \in \mathbb{R}^{|G|}$, where $|G|$ is the number of genes, and the entries of $\mathbf{M}_j$ contain the gene counts observed in R_j. This yields a set of "ST tokens", each spatially aligned to an image token and characterized by the local gene expression profile $\mathbf{M}_j$. Importantly, this approach circumvents the need for explicit cell segmentation of the H&E and ST images, a tedious and error-prone process [13] that would be otherwise required to assign transcripts to individual cells and link them back to their counterparts in the H&E image. Each token's gene expression profile $\mathbf{M}_j$ is encoded by a dedicated transcript encoder which plays a similar role as the PatchEmbed layer in the ViT. In our experiments, we mainly use GeneFormer [6, 14] models, but any gene expression encoder can be plugged into this framework with minimal changes.

Modality fusion To fully utilize the flexibility of ViT to handle sequences of variable lengths, we fuse the modalities by expanding the sequence of the tokens. The two sets of tokens are concatenated into one longer sequence after projecting to d dimensions and adding positional encodings. For example, a 14×14 patch grid could yield 196 image tokens + 196 transcript tokens + 1 *cls* token = 393 tokens. For comparison, we also considered a different fusion strategy, where we combine the modalities at each token-patch position in the feature dimension without sequence expansion, i.e., the image and transcriptomic token embeddings are either averaged or concatenated in the feature dimension at each token-patch position. After fusion, the combined token sequence is fed into the self-attention layers of the ViT. The *cls* token attends to both H&E and ST tokens, thus capturing a joint tissue representation. Importantly, our design is *modality-flexible* as the tokens from the two modalities are treated as interoperable tokens: with only H&E tokens available, the model reduces to the original ViT; with only ST tokens present, the model provides a novel way to aggregate ST data.

Training Strategy To illustrate the efficacy of the proposed model architecture, we perform supervised learning using our model on the disease state classification task in HEST-1k dataset [15] (details in section 3). The sequence of embeddings after the transformer layers are pooled together to create image-level embeddings which are fed into a linear classification head for downstream tasks. Different pooling strategies, such as averaging, or using cls token can be applied. During training, both the ViT backbone and the gene expression encoder are kept frozen, and only the projection layer from the transcriptomic embedding to the vision embedding space, and the classification head are trained, as shown in figure 1 (more implementation details in the Appendix A.2). The supervised approach could accommodate any downstream tasks, and the small number of trainable parameters reduces the risk of overfitting.

3 Results

Data We tested our model on the HEST-1k dataset [15], a publicly available collection of ST profiles with corresponding WSIs and metadata. We focused on the Xenium subset of HEST-1k that contains 59 pairs of Xenium and H&E images, covering a diverse collection of human samples from 14 organs and 18 tissue types, further labeled by disease state (18 diseased, 28 cancer, 13 healthy, details in Appendix Figure 3). The experiments were conducted at a patch level, with the slide-level disease states propagated to be patch-level labels. To evaluate the model's performance, we created 4-fold stratified train/test splits based on the sample level, with an average train/test ratio of 75/25, corresponding to ∼300,000 and ∼100,000 patches.

Fusing modalities achieves higher performance in disease state prediction To evaluate the effect of fusing the H&E and Xenium images, we compared the performances of uni- and multi-modal models trained with different choices of the fusion and pooling strategies. All experiments here used the ViT-B14 from the Midnight series as the vision encoder and "gf-6L-30M-i2048" GeneFormer as the transcripts encoder. All results across all splits are given in Table 1 in terms of macro-accuracy (and Table 2 in Appendix for F1 scores), and additionally in Figure 2 as the difference to the performance of ResNet18 [16] trained on H&E images, used as a baseline. Our preliminary results indicate that, across all 4 splits, different versions of the multimodal fusion model were always the top performing, with the sequence expansion and average pooling configuration ranking first in 2 out of the 4 splits. We also see large variations across the 4 splits, most likely due to the small sample size of the dataset: while in most cases the top scoring models exceeded a macro-accuracy of 0.85, in Split 1 almost all models struggled to exceed a macro-accuracy of 0.7. Interestingly, the "expr-only-image"

Table 1: Macro-accuracy of disease state prediction by various models across four splits. **Bold** is best.

Model	Modality		Fusion	Pooling	Macro-accuracy (↑)			
	H&E	ST			Split 1	Split 2	Split 3	Split 4
fmx-concat-avg	✓	✓	concat	average	0.61	**0.89**	**0.86**	0.65
fmx-concat-token	✓	✓	concat	cls token	0.65	0.66	0.71	**0.94**
fmx-add-avg	✓	✓	sum	average	0.63	0.65	0.67	0.71
fmx-add-token	✓	✓	sum	cls token	**0.67**	0.65	0.67	0.70
expr-only-image		✓	NA	NA	0.60	0.72	0.78	0.88
expr-only-token		✓	NA	cls token	0.49	0.57	0.56	0.57
vision-only-avg	✓		NA	average	0.59	0.54	0.66	0.59
vision-only-token	✓		NA	cls token	0.62	0.56	0.69	0.62
ResNet-18	✓		NA	NA	0.53	0.57	0.66	0.52

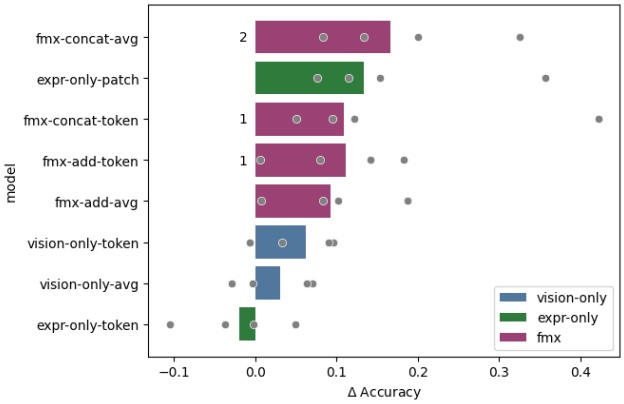

Figure 2: Performance of various uni- and multi-modal models, in comparison to ResNet18 in terms of macro-accuracy. For each data split and each model configuration, we report Δ Accuracy, defined as the difference of the macro-accuracy of each model to the one of ResNet18. Bars indicate the median of Δ Accuracy across 4 splits. Numbers under the bars indicate in how many of the 4 splits the model ranked first in terms of absolute macro-accuracy.

variant ranked second in terms of median macro-accuracy, outperforming the "expr-only-token" variant (see section A.2 for more details on the differences between the two variants). We suspect that the unexpected low performance of the "expr-only-token" variant is largely due to the sparsity of the transcripts at token level, i.e. no or fewer transcripts at token level thus higher noise; as well as the distribution shift of the token level gene expression profile from the cell level profiles with which the GeneFormer models were trained. Finally, both vision-only models achieved a very low performance regardless of the pooling strategy, indicating that morphology alone is not enough for the task at hand.

4 Discussion and Future Work

Although our results demonstrate that token-level fusion of H&E and ST consistently improves disease state prediction compared to unimodal models, the variability across folds highlights the limitations imposed by small sample sizes and heterogeneous data sources, suggesting that more robust evaluation requires larger benchmarks. Although our results are encouraging and establish the potential of fusing modalities early, they are still preliminary, and we are currently working on a number of additional baselines and extensions, including: (i) testing more FM backbone models, (ii) pretraining the transcript encoder to overcome the potential distributional shift, (iii) training the model on additional tasks (e.g., tissue type prediction), (iv) scaling to larger datasets to reduce variance and improve generalization. As future work, we aim to train our model in a self-supervised fashion, e.g. by continued pretraining of a vision-only encoder on a dataset with matched H&E and ST data.

References

[1] Aditya Kashyap, Maria Anna Rapsomaniki, Vesna Barros, Anna Fomitcheva-Khartchenko, Adriano Luca Martinelli, Antonio Foncubierta Rodriguez, Maria Gabrani, Michal Rosen-Zvi, and Govind Kaigala. Quantification of tumor heterogeneity: from data acquisition to metric generation. *Trends in Biotechnology*, 40(6):647–676, 2022.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

[3] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

[4] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A. Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, October 2024. Publisher: Nature Publishing Group.

[5] Mikhail Karasikov, Joost van Doorn, Nicolas Känzig, Melis Erdal Cesur, Hugo Mark Horlings, Robert Berke, Fei Tang, and Sebastian Otálora. Training state-of-the-art pathology foundation models with orders of magnitude less data, 2025.

[6] C V Theodoris, L Xiao, A Chopra, M D Chaffin, Z R Al Sayed, M C Hill, H Mantineo, E Brydon, Z Zeng, X S Liu, and P T Ellinor. Transfer learning enables predictions in network biology. *Nature*, 2023.

[7] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, pages 1–11, February 2024. Publisher: Nature Publishing Group.

[8] Dario Bressan, Giorgia Battistoni, and Gregory J. Hannon. The dawn of spatial omics. *Science*, 381(6657):eabq4964, August 2023. Publisher: American Association for the Advancement of Science.

[9] Amanda Janesick, Robert Shelansky, Andrew D. Gottscho, Florian Wagner, Stephen R. Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A. Morrison, Michelli F. Oliveira, Jordan T. Sicherman, Andrew Kohlway, Jawad Abousoud, Tingsheng Yu Drennon, Seayar H. Mohabbat, and Sarah E. B. Taylor. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1):8353, December 2023. Number: 1 Publisher: Nature Publishing Group.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[11] Guillaume Jaume, Lukas Oldenburg, Anurag Vaidya, Richard J. Chen, Drew F. K. Williamson, Thomas Peeters, Andrew H. Song, and Faisal Mahmood. Transcriptomics-guided Slide Representation Learning in Computational Pathology, May 2024. arXiv:2405.11618 [cs].

[12] Weiqing Chen, Pengzhi Zhang, Tu N. Tran, Yiwei Xiao, Shengyu Li, Vrutant V. Shah, Hao Cheng, Kristopher W. Brannan, Keith Youker, Li Lai, Longhou Fang, Yu Yang, Nhat-Tu Le, Jun-ichi Abe, Shu-Hsia Chen, Qin Ma, Ken Chen, Qianqian Song, John P. Cooke, and Guangyu Wang. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, 22(7):1568–1582, July 2025. Publisher: Nature Publishing Group.

[13] Mariia Bilous, Daria Buszta, Jonathan Bac, Senbai Kang, Yixing Dong, Stephanie Tissot, Sylvie Andre, Marina Alexandre-Gaveta, Christel Voize, Solange Peters, Krisztian Homicsko, and Raphael Gottardo. From Transcripts to Cells: Dissecting Sensitivity, Signal Contamination, and Specificity in Xenium Spatial Transcriptomics, April 2025. Pages: 2025.04.23.649965 Section: New Results.

[14] H Chen, M S Venkatesh, J Gomez Ortega, S V Mahesh, T Nandi, R Madduri, K Pelka†, and C V Theodoris. Quantized multi-task learning for context-specific representations of gene network dynamics. *bioRxiv*, 2024.

[15] Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Perez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems*, December 2024.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

A Appendix

A.1 HEST-1k Xenium subset

Based on the alignment between the H&E slides and transcripts data made available by the HEST-1k dataset, we cropped the H&E slides to the bounding boxes covering all transcripts data and extracted patches of 256 x 256 pixels at a resolution of 0.25 micron per pixel. Patches with a foreground area less than 25% based on the tissue segmentation in HEST-1K data were dropped.

Although the disease type label is at a slide level, the experiments were conducted at a patch level, and the slide level labels were simply propagated to be the patch level labels. To evaluate the model performance, we created 4-fold stratified train/test splits with a ratio of 75/25 based on the sample level. On average in each split there are $\sim$300,000 and $\sim$100,000 patches.

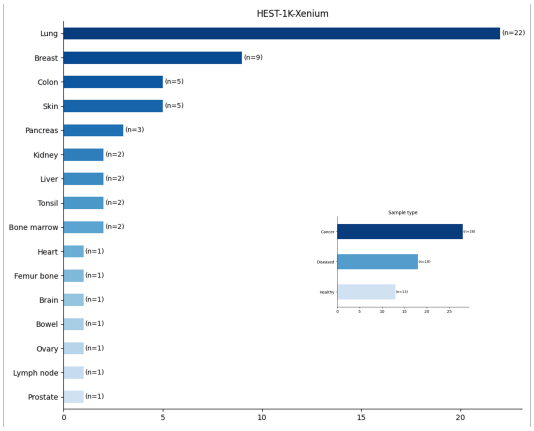

Figure 3: Overview of Xenium subset in HEST-1K: 59 Homo Sapien samples with H&E slides and corresponding transcripts data.

A.2 Implementation details

For the vision encoder, we experimented with the "Midnight-12k"[5] model of size ViT-g14 and a smaller ViT-B14 model from the same series. For the gene expression encoder, we used the "gf-6L-30M-i2048" and "gf-18L-316M-i4096" version of the GeneFormer[6, 14]. The model was trained with the AdamW optimizer using a learning rate of 0.00001, weight decay of 0.04, and a global batch size of 256 when transcripts are included in the input and 1024 for vision only inputs. Further ablation studies on hyperparameters, model sizes are still work in progress.

The model and training framework were implemented using PyTorch and pytorch-lightning libraries. The experiments were performed on two nvidia H200-80GB GPUs. Each model configuration was trained for maximally 15 epochs or 24 hours, whichever comes first.

Aggregating transcripts at image vs. token level For the configurations with transcripts only inputs ("expr-only-image" and "expr-only-token" in table 1), we compared two strategies to aggregate the transcripts:

- "expr-only-token": as described in section 2, the transcripts are aggregated within each token, individually encoded by GeneFormer and then average pooled to obtain a patch-level representation. A token of 16x16 pixels with a resolution of $0.25\mu m$ per pixel (mpp), corresponds to an area of $16\mu m^2$, which is smaller than the typical size of a cell which range between $10 - 20\mu m$ in diameter or around $150\mu m^2$. Thus, the gene expression profile for each token could be out-of-distribution with respect to the training data seen by GeneFormer, which was pretrained on single-cell transcriptomics data. This would reduce the quality of the token embeddings and explain the markedly lower performance to the "expr-only-image" version.

- "expr-only-image": all the transcripts in the whole image area are aggregated and then encoded with the gene expression encoder to obtain the patch-level embedding. An image of size 256×256 pixels at a resolution of 0.25 mpp spans an area of 64×64 μm, which is considerably larger than individual cell sizes. However, since GeneFormer relies only on the ranked gene expression values, aggregating expression over such regions reduces noise and produces profiles more consistent with the type of data GeneFormer was trained on, thereby avoiding performance loss.

A.3 Additional results

Table 2: F1 score of disease state prediction by various models across four splits. **Bold** is best.

Model	**Modality**		**Fusion**	**Pooling**	**F1 score (↑)**			
	H&E	ST			Split 1	Split 2	Split 3	Split 4
fmx-concat-avg	✓	✓	concat	average	0.67	**0.90**	**0.87**	0.60
fmx-concat-token	✓	✓	concat	cls token	0.71	0.70	0.74	**0.90**
fmx-add-avg	✓	✓	sum	average	0.69	0.67	0.68	0.68
fmx-add-token	✓	✓	sum	cls token	**0.73**	0.66	0.68	0.67
expr-only-image		✓	NA	NA	0.67	0.76	0.82	0.85
expr-only-token		✓	NA	cls token	0.50	0.57	0.56	0.57
vision-only-avg	✓		NA	average	0.58	0.55	0.66	0.58
vision-only-token	✓		NA	cls token	0.62	0.57	0.70	0.61
ResNet-18	✓		NA	NA	0.55	0.58	0.64	0.51

A.4 Model and data licenses

The model licenses for the models in this work are as follows:

- **ResNet-18:** Qualcomm® license can be found at here: `https://qaihub-public-assets.s3.us-west-2.amazonaws.com/qai-hub-models/Qualcomm+AI+Hub+Proprietary+License.pdf`
- **Vision encoder "Midnight-12k":** permissive MIT-license
- **Expression-only "gf-6L-30M-i2048" and "gf-18L-316M-i4096":** Apache-2.0
- **HEST-1k data:** cc-by-nc-sa-4.0

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our results support the claims in the introduction and abstract. As this is a work in progress paper, we have also clarified across the paper that the results are preliminary.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: As this is an work in progress paper, we discuss current limitations and future work to overcome these limitations in the "Discussion and Future Work" section of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: No theoretical results are presented in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Figure 1, the Methods section and additional details in the Appendix describe the model architecture, dataset, encoding strategy and additional steps needed to recreate the results presented in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The dataset is publicly available and can be downloaded following the instructions here: `https://huggingface.co/datasets/MahmoodLab/hest`. The code is not yet public as this is work in progress.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Data splits are described in the Data subsection in the Results section. Model training details including optimizer, hyperparameters are described in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Figure 2 shows the median results, to show the single data points to show the statistical variation around this mean. Otherwise no summary statistics are reported.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The 'Implementation details' in the appendix describe the computational resources used for all experiments?.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform to the ethical guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This paper is a work in progress paper and at this stage it is too early to discuss how and whether this work would impact the society in for example health care applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models and data that are not original are properly cited and can be found in the references. The corresponding licenses are also mentioned in Appendix section A.4.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Although we present a new method in this paper, due to the early stage of this research it is not yet published as a tool for users. Therefor, no new assets are released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.