MulTaBench: Benchmarking Multimodal Tabular Learning with Text and Image

Anonymous Authors[1]

Abstract

Tabular Foundation Models have recently established the state of the art in supervised tabular learning. However, they lack native support for unstructured modalities such as text and image, and rely on frozen, pretrained embeddings to process them. We show that tuning the embeddings to the task improves performance on established Multimodal Tabular benchmarks. We introduce MulTaBench, a benchmark of 40 datasets, split equally between image-tabular and text-tabular tasks. We focus on predictive tasks where the modalities provide complementary predictive signal, and where generic embeddings lose critical information, necessitating Target-Aware Representations that are aligned with the task. We demonstrate that the gains from target-aware representation tuning generalize across both text and image modalities, several tabular learners, encoder scales, and embedding dimensions. MulTaBench constitutes the largest image-tabular benchmarking effort to date, enabling the research of novel architectures which incorporate target-aware representations, paving the way for the development of novel Multimodal Tabular Foundation Models.

1. Introduction

Tabular Foundation Models (TFMs) (Van Breugel & Van Der Schaar, 2024; Hollmann et al., 2022; 2025) have recently emerged as the state of the art (SOTA) for supervised tabular learning (Erickson et al., 2025). However, the best-performing TFMs (Grinsztajn et al., 2026; Qu et al., 2026) are trained exclusively on structured numerical data, making them fundamentally unimodal: unstructured inputs must be preprocessed via external embedding models, with no unified support for modalities such as text and image.

Yet, in many high-impact domains, tabular problems are multimodal: e-commerce listings, social media feeds, and medical health records combine image and text with numerical features. While early work has begun extending TFMs to integrate text (Arazi et al., 2025; Spinaci et al., 2025), these extensions often compromise the model's core tabular performance, and inherent support for visual modalities remains entirely absent. One might turn to Large Language and Vision-Language Models (LLMs/VLMs), which natively process unstructured inputs, but they are not suited for the inductive biases of tabular data; specifically, they are unoptimized for the relational structure (Fang et al., 2024) and are suboptimal for numerical features (Van Breugel & Van Der Schaar, 2024). Addressing these limitations requires architectures that combine the numerical precision of TFMs while maintaining the rich input handling of multimodal foundation models. However, evaluating such a unified approach is difficult because the diverse nature of tasks within Multimodal Tabular Learning (MMTL) is not yet fully characterized; existing benchmarks (Shi et al., 2021; Lu et al., 2023; Kim et al., 2024; Tang et al., 2024b; Mráz et al., 2025) primarily highlight the coexistence of modalities, unintentionally grouping together problems that require fundamentally different modeling solutions.

To characterize these problems, we observe that tabular models require inputs to be represented as feature columns, so high-dimensional images and texts must be compressed into compact representations. Consequently, embeddings act as lossy summaries, as they capture only a fraction of the raw input's information by design (Weller et al., 2025). In order to generalize well, pretrained embedding models are optimized for broad semantic content, such as distinguishing an X-ray from a mammogram, at the expense of fine-grained details like precise size estimations or localized anomalies (Pantazopoulos et al., 2024; Li et al., 2025). While this compression is effective for global semantic mapping, it fails to preserve the specialized signals required for fine-grained MMTL tasks. We thus advocate for the need for Target-Aware Representations (TAR): embeddings that are tuned to the target and, ideally, to the other modalities.

Consider, for example, the task of pneumonia detection from a patient record combining age and smoking status with chest X-ray images. We argue that to study MMTL, a dataset should satisfy two properties: (1) *Joint Signal*, where each modality provides complementary information

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

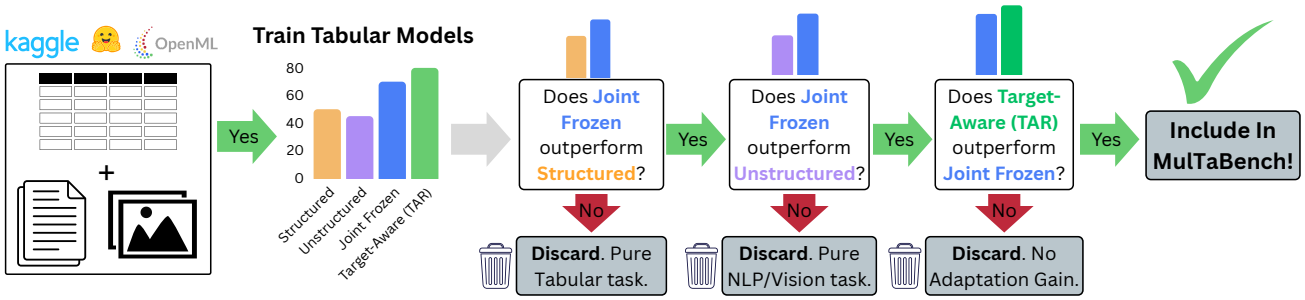

Figure 1. The MulTaBench Curation Pipeline. Datasets are included if they require *Joint Signal* and *Task-awareness*.

that contributes to the overall predictive performance, and (2) *Task-awareness*, where task-agnostic representations fail to capture the details required for a given objective. In our example, both the X-ray and the clinical profile offer unique, complementary information, and steering the image embedding to detect subtle signs of inflammation in the lungs should improve diagnostic accuracy.

To translate these theoretical properties into a measurable test, we develop an algorithmic pipeline that quantifies whether a dataset complies with the aforementioned requirements. This approach approximates these properties by evaluating each task across a broad suite of tabular learners, ranging from light GBDTs to SOTA TFMs. To evaluate for *Joint Signal*, we demand a performance drop when either modality is removed, verifying that each input strengthens the predictive power. For *Task-awareness*, we finetune the encoder's last 3 layers with LoRA (Hu et al., 2021) on the prediction target as a preprocessing step, and we expect these representations to outperform frozen ones when passed to tabular models. Crucially, our experiments confirm that TAR outperform frozen embeddings across established MMTL benchmarks; however, we find that the magnitude of these gains is highly dataset-dependent, suggesting they represent distinct classes of MMTL tasks.

Building on this framework, we introduce **MulTaBench**, a benchmark of 40 datasets balanced between image-tabular and text-tabular tasks, as well as classification and regression objectives. To ensure a comprehensive evaluation, the benchmark incorporates a wide range of sample sizes and feature counts, while spanning a diverse set of domains to capture the heterogeneity of real-world multimodal tabular data. MulTaBench represents the largest image-tabular benchmarking effort to date, and the first MMTL benchmark to explicitly prioritize datasets requiring task-aware representations. Demonstrating the robustness of our curation criteria, we show that the gains from target-aware tuning generalize consistently across a diverse suite of independent tabular learners, encoder scales, and embedding dimensions. These findings suggest that designing novel architectures which contextualize the representations of unstructured modalities can push the boundaries of MMTL, and we believe that MulTaBench would be instrumental for developing true Multimodal TFMs. See Appendix A for extended introduction, and Appendix B for Related Work.

2. Benchmarking MMTL

2.1. Desiderata for Multimodal Tabular datasets

Joint Signal. Following the principle in Mráz et al. (2025), we require each modality to carry independent signal about the target, so the joint predictive performance exceeds the union of unimodal performances. In the pneumonia case, the X-ray encodes spatial lung patterns, while age and smoking status convey clinical risk factors that provide information invisible in pixels. This criterion could optionally capture cross-modal interactions, where one modality might only become discriminative once conditioned on the other. For instance, increased reticular markings may signal acute infection in non-smokers, yet merely represent baseline chronic changes in a long-term smoker; the visual feature only becomes discriminative when conditioned on the tabular history. A modality can fail this criterion if it carries no signal, or if its signal is already captured by another modality and thus provides no predictive gain.

Task-awareness A task exhibits *Task-awareness* when the predictive signal is latent in the raw input at a level of granularity that differs from the modality's global semantic meaning. Because general-purpose encoders are optimized to preserve high-level properties while discarding low-level variance, such as exact wording (Weller et al., 2025) or fine-grained spatial textures (Pantazopoulos et al., 2024), they often discard the specific nuances required for MMTL. Recovering this signal necessitates TAR, which steer the representation to focus on the details relevant to the specific target. In our pneumonia example, a generic model might identify the scan's global anatomy, whereas TAR would preserve the tiny visual patterns in the lung tissue that are key for diagnosis. Conversely, a task lacks *Task-awareness*

if the predictive signal is coarse enough to be captured by task-agnostic embeddings; for instance, if the objective is simply to categorize the scan type rather than identify a specific pathology, TAR would provide no significant advantage.

2.2. The Curation Pipeline

To bridge the gap between the theoretical desiderata and the empirical curation, we establish an evaluation protocol based on 4 experimental conditions, as summarized in Figure 1 and Table 1 in Appendix C. The conditions vary by the features included and the specific representation of the unstructured modalities. Our approach intentionally entangles task properties with algorithmic solutions in order to isolate datasets that align with our criteria and that current models struggle with. Embeddings are extracted using *e5-v2-small* (Wang et al., 2024) for texts and *DINO-v3-small* (Siméoni et al., 2025) for images. To implement our proposed TAR condition, we finetune the last 3 layers on the prediction target using LoRA. Crucially, this adaptation is performed as a specialized preprocessing step without the structured features and shared across learners. Representations are down-projected with PCA to a dimension of 30, to ensure computational efficiency. We employ 5 diverse tabular learners. For each candidate dataset, we evaluate every model in each condition over 5 random seeds, subsampling up to 10,000 examples per run for cost-effectiveness. Our metric is AUC for classification tasks and R^2 for regression tasks.

Acceptance Criteria. To pass the curation filter, a dataset should satisfy two conditions across at least 3 out of 5 learners: (1) For *Joint Signal*, performance over the *Joint Frozen* condition should be higher than both *Unimodal Structured* and *Unimodal Unstructured* variants. This ensures that the unstructured modality is relevant, while also prevents the dataset from collapsing into a pure Natural Language Processing or Computer Vision task; and (2) For *Task-awareness* we require that the *Joint TAR* condition will improve performance over the *Joint Frozen* condition, isolating the gain from representation tuning. Figure 4 in the Appendix illustrates the protocol over concrete examples, and Appendix C provides a formal and precise definition of the acceptance criteria, and details of the curation setup.

3. MulTaBench

MulTaBench is composed of 40 datasets split equally between image-tabular and text-tabular while balancing between regression and classification tasks, all satisfying our curation pipeline established in §2. While the text-tabular subset is derived exclusively from existing benchmarks, the image-tabular subset is curated and collected from public datasets. A comprehensive summary of the benchmark is provided in Appendix D.

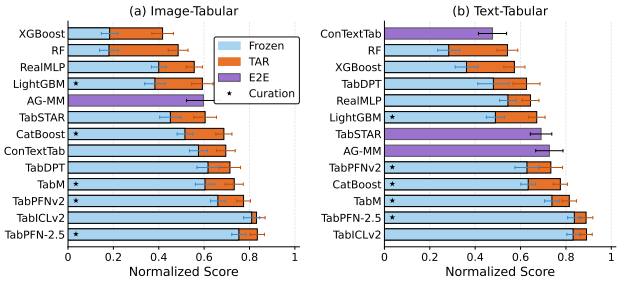

Figure 2. All learners gain from Target-Aware Representations.

Text-Tabular Curation. To evaluate existing text-tabular benchmarks (Shi et al., 2021; Grinsztajn et al., 2023; Kim et al., 2024; Mráz et al., 2025), we aggregate all their 56 unique datasets and subject them to our 4 experimental conditions. In Figure 6 in the Appendix, we compare *Joint TAR* and *Joint Frozen* across all datasets, finding that TAR consistently outperforms frozen embeddings for all learners, highlighting the limitations of using fixed representations. With the results in hand, we apply our curation pipeline and find out that approximately 23% of the datasets fail the *Joint Signal* criterion; of the remaining datasets, 36% do not pass the *Task-awareness* criterion, leaving 41% that pass both. From these, we subsample 20 datasets to match the size of the image-tabular subset. Our acceptance rate shows that while our requirements are common enough, they do not constitute the primary focus of standard text-tabular research. Without this distinction, existing benchmarks lack the focus to research target-awareness in MMTL.

Image-Tabular Curation. We collect candidate datasets from existing literature (Lu et al., 2023; Tang et al., 2024b; Luo et al., 2025b; Kim et al., 2025b), identifying a shared pool of 16 unique valid datasets, from which only 5 meet our criteria (31%), a proportion comparable to the text-tabular subset. We then manually curate additional datasets from Kaggle which pass our pipeline, eventually creating the largest image-tabular benchmark to this date with 20 datasets. In the process, we encountered significant challenges, detailed in Appendix F.

4. Robustness Analysis

While our curation pipeline identifies datasets with high multimodal potential, it is crucial to verify that these properties remain consistent across different modeling choices.

New Tabular Learners. Since model ranking suffers from selection bias favoring the curation models, our objective is not to establish the SOTA, but to provide a useful tool for the development of future multimodal architectures. We supplement the original learners with 5 additional ones,

while also including 3 "end-to-end" (E2E) models which natively processes texts or images. Figure 2 shows model performance on both MulTaBench subsets. Target-aware embeddings consistently outperform frozen embeddings across all new models and modalities. While this gain is expected for the curation models, its generalization to all the other models provides an indication to the usefulness of our benchmark for MMTL research.

Embedding Model Scale. So far, text and image were represented using *e5-v2-small* and *DiNO-v3-small*. Since the dimension of these embeddings is 384, one potential limitation may be that they are too small. We thus repeat the curation experiments using the *Large* variants of the models, which have approximately 10 times more parameters, and a final dimension of 1024. Figure 10 in the Appendix shows that while a larger embedding model improves downstream performance, TAR significantly outperforms frozen embeddings even at the larger scale. In fact, we even observe that the *TAR Small* variant is better than *Frozen Large*; this indicates that increased representational capacity does not guarantee that target-relevant signals are retained in the final representation, and tuning is still required.

Embedding Dimension. To this point, our analysis has assumed a fixed embedding size of 30 PCA components, following standard practice (Grinsztajn et al., 2023; Arazi et al., 2025). This dimensionality reduction helps prevent overfitting and ensures computational efficiency by reducing memory requirements. However, this raises the question: is TAR really surfacing information which was missing in the original representations, or is the observed gain an artifact of the compression? In Appendix H.5, we show that representation tuning remains effective across 15 and 60 dimensions, and even when removing PCA completely.

Qualitative Analysis. Figure 3 and Appendix I illustrate how target-aware adaptation reshapes the encoder's focus across 4 MulTaBench datasets. In *CheXpert*, attention shifts from arbitrary anatomical borders toward the right lower lung and optic disc, respectively. Similarly, focus in *Celebs* moves from peripheral accessories to core facial features. These examples demonstrate that contextualization enables the encoder to surface specific details that are otherwise lost in task-agnostic representations.

5. Towards Multimodal TFMs

Our analysis of MulTaBench reveals a significant gap between current tabular learners and the demands of MMTL tasks, as existing architectures cannot jointly tune unstructured representations for the target label. In this section, we discuss the potential trajectory of future Multimodal TFMs. Our vision builds upon the framework proposed

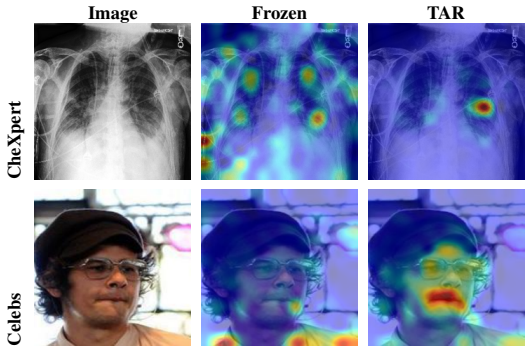

Figure 3. Comparison of frozen and Target-Aware Representations.

by Van Breugel & Van Der Schaar (2024). Their position piece identifies TFMs as a research priority and defines 4 core desiderata to guide their development: (D1) handling *mixed-type columns*, such as numbers, categories and dates, (D2) enabling *cross-dataset modeling*, (D3) leveraging *textual context* and metadata, such as column names, and (D4) maintaining *equivariance to column order*. We expand their definition by suggesting (D5) *Target-Aware Multimodal Tabular Learning*; text and image embeddings should be target-aware.

While PFNs have revolutionized structured learning, they are primarily designed for modalities where raw inputs already contain highly compressed signals. Initial efforts attempting to couple PFNs with multimodal encoders (Luo et al., 2025b; Kim et al., 2025b) have struggled to unlock TAR without violating the core ICL premise of avoiding parameter updates. In contrast, joint modeling approaches such as AutoGluon-Multimodal and TabSTAR utilize finetuning to achieve target-awareness, yet this introduces significant practical challenges. Finetuning historically complicates tabular learning by increasing overfitting risks, particularly on small-to-medium datasets, and imposing substantial computational overhead as data, model and embedding scales grow. This burden increases further when using HPO or standard practices like cross-validation and ensembling, as these methods require repeating the expensive finetuning process multiple times to find the best parameters and prevent data leakage across splits.

To summarize, we argue that none of the current architectures are optimal for MMTL, and that the leading paradigms complement each other. MulTaBench enables their development by isolating the datasets that explicitly demand task-specific representations. While proposing a solution is out of this work's scope, we believe that the optimal architecture should take the best of both worlds. An ideal model should bring the contextualization benefits of TAR while preserving the robustness and latency of ICL. We hope that the existence of MulTaBench will enable the research of such models (see Appendix J for further discussion).

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a Visual Language Model for Few-Shot Learning. October 2022. URL https://openreview.net/forum?id=EbMuimAbPbs.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, Santiago, Chile, December 2015. IEEE. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.279. URL http://ieeexplore.ieee.org/document/7410636/.

Arazi, A., Shapira, E., and Reichart, R. TabSTAR: A Tabular Foundation Model for Tabular Data with Text Fields. October 2025. URL https://openreview.net/forum?id=FrXHdcTEzE.

Badian, Y., Ophir, Y., Tikochinski, R., Calderon, N., Klomek, A. B., Fruchter, E., and Reichart, R. Social media images can predict suicide risk using interpretable large language-vision models. *J Clin Psychiatry*, 85(1): 50516.

Bordt, S., Nori, H., Rodrigues, V., Nushi, B., and Caruana, R. Elephants Never Forget: Memorization and Learning of Tabular Data in Large Language Models. *First Conference on Language Modeling*, 2024.

Bouadi, M., Seth, P., Tanna, A., and Sankarapu, V. K. Orion-MSP: Multi-Scale Sparse Attention for Tabular In-Context Learning, November 2025. URL http://arxiv.org/abs/2511.02818. arXiv:2511.02818 [cs].

Brahmavar, S. B., Liu, Q., Li, Y., and Oliva, J. Task Expansion and Cross Refinement for Open-World Conditional Modeling, March 2026. URL http://arxiv.org/abs/2603.13308. arXiv:2603.13308 [cs].

Breiman, L. Random Forests. *Machine Learning*, 45(1): 5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Baraldi, L., Cornia, M., and Cucchiara, R. The Revolution of Multimodal Large Language Models: A Survey. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13590–13618, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.807. URL https://aclanthology.org/2024.findings-acl.807/.

Cao, B., Chen, K., Maninis, K.-K., Chen, K., Karpur, A., Xia, Y., Dua, S., Dabral, T., Han, G., Han, B., Ainslie, J., Bewley, A., Jacob, M., Wagner, R., Ramos, W., Choromanski, K., Seyedhosseini, M., Zhou, H., and Araujo, A. TIPSv2: Advancing Vision-Language Pretraining with Enhanced Patch-Text Alignment, April 2026. URL http://arxiv.org/abs/2604.12012. arXiv:2604.12012 [cs].

Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, August 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL https://dl.acm.org/doi/10.1145/2939672.2939785.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., Jiang, N.-J., et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, December 2025. URL http://arxiv.org/abs/2507.06261. arXiv:2507.06261 [cs].

Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L. A., Wilson, K. T., Landman, B. A., and Huo, Y. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Progress in Biomedical Engineering*, 5(2):022001, April 2023. ISSN 2516-1091. doi: 10.1088/2516-1091/acc2fe. URL https://doi.org/10.1088/2516-1091/acc2fe.

Das, R., Ahmed, W., Sharma, K., Hardey, M., Dwivedi, Y. K., Zhang, Z., Apostolidis, C., and Filieri, R. Towards the development of an explainable e-commerce fake review index: An attribute analytics approach. *European Journal of Operational Research*, 317(2):382–400, 2024.

Du, S., Zheng, S., Wang, Y., Bai, W., O'Regan, D. P., and Qin, C. TIP: Tabular-Image Pre-training for Multimodal Classification with Incomplete Data. In Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., and Varol, G. (eds.), *Computer Vision – ECCV 2024*, volume 15073, pp. 478–496. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-72632-3 978-3-031-72633-0. doi: 10.1007/978-3-031-72633-0_27. URL https://link.springer.com/10.1007/978-3-031-72633-0_27. Series Title: Lecture Notes in Computer Science.

Duenias, D., Nichyporuk, B., Arbel, T., and Raviv, T. R. HyperFusion: A Hypernetwork Approach to Multimodal Integration of Tabular and Medical Imaging Data for Predictive Modeling. *Medical Image Analysis*, 102:103503, May 2025. ISSN 13618415. doi: 10.1016/j.media.2025.103503. URL http://arxiv.org/abs/2403.13319. arXiv:2403.13319 [cs].

Ebrahimi, S., Arik, S. O., Dong, Y., and Pfister, T. LANISTR: Multimodal Learning from Structured and Unstructured Data, April 2024. URL http://arxiv.org/abs/2305.16556. arXiv:2305.16556 [cs].

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., and Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization, February 2025. URL http://arxiv.org/abs/2404.16130. arXiv:2404.16130 [cs].

Eggert, G., Huo, K., Biven, M., and Waugh, J. TabLib: A Dataset of 627M Tables with Context, October 2023. URL http://arxiv.org/abs/2310.07875. arXiv:2310.07875 [cs].

Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Desai, P. M., Salinas, a. D., and Hutter, F. TabArena: A Living Benchmark for Machine Learning on Tabular Data, June 2025. URL http://arxiv.org/abs/2506.16791. arXiv:2506.16791 [cs].

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., and Li, Q. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 6491–6501, New York, NY, USA, August 2024. Association for Computing Machinery. ISBN 979-8-4007-0490-1. doi: 10.1145/3637528.3671470. URL https://dl.acm.org/doi/10.1145/3637528.3671470.

Fang, X., Xu, W., Tan, F. A., Hu, Z., Zhang, J., Qi, Y., Sengamedu, S. H., and Faloutsos, C. Large Language Models (LLMs) on Tabular Data: Prediction, Generation, and Understanding - A Survey. *Transactions on Machine Learning Research*, March 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=IZnrCGF9WI.

Fu, Y., Zhao, Y., Zeng, Z., Chen, C., and Jin, Y. Unleashing the Power of Image-Tabular Self-Supervised Learning via Breaking Cross-Tabular Barriers, December 2025. URL http://arxiv.org/abs/2512.14026. arXiv:2512.14026 [cs].

Ganz, R., Kittenplon, Y., Aberdam, A., Avraham, E. B., Nuriel, O., Mazor, S., and Litman, R. Question Aware Vision Transformer for Multimodal Reasoning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13861–13871, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.01315. URL https://ieeexplore.ieee.org/document/10655638/.

Gardner, J., Perdomo, J. C., and Schmidt, L. Large Scale Transfer Learning for Tabular Data via Language Modeling. *Advances in Neural Information Processing Systems*, 37:45155–45205, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4fd5cfd2e31bebbccfa5ffa354c04bdc-Abstract-Conference.html.

Garg, A., Ali, M., Hollmann, N., Purucker, L., Müller, S., and Hutter, F. Real-TabPFN: Improving Tabular Foundation Models via Continued Pre-training With Real-World Data. June 2025. URL https://openreview.net/forum?id=BtEiqKsIMw.

Gisserot-Boukhlef, H., Boizard, N., Faysse, M., Alves, D. M., Malherbe, E., Martins, A., Hudelot, C., and Colombo, P. Should We Still Pretrain Encoders with Masked Language Modeling? October 2025. URL https://openreview.net/forum?id=jpz7e3jhRq.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. Revisiting deep learning models for tabular data. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, pp. 18932–18943, Red Hook, NY, USA, December 2021. Curran Associates Inc. ISBN 978-1-7138-4539-3.

Gorla, A. and Puduppully, R. The Illusion of Generalization: Re-examining Tabular Language Model Evaluation, February 2026. URL http://arxiv.org/abs/2602.04031. arXiv:2602.04031 [cs] version: 1.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35:507–520, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html.

Grinsztajn, L., Oyallon, E., Kim, M. J., and Varoquaux, G. Vectorizing string entries for data processing on tables: when are larger language models better?, December 2023. URL http://arxiv.org/abs/2312.09634. arXiv:2312.09634 [stat].

Grinsztajn, L., Flöge, K., Key, O., Birkel, F., Jund, P., Roof, B., Jäger, B., Safaric, D., Alessi, S., Hayler, A., Manium, M., Yu, R., Jablonski, F., Hoo, S. B., Garg, A., Robertson, J., Bühler, M., Moroshan, V., Purucker, L., Cornu, C., Wehrhahn, L. C., Bonetto, A., Schölkopf, B., Gambhir, S., Hollmann, N., and Hutter, F. TabPFN-2.5: Advancing the State of the Art in Tabular Foundation Models, February 2026. URL http://arxiv.org/abs/2511.08667. arXiv:2511.08667 [cs].

Hager, P., Menten, M. J., and Rueckert, D. Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data, March 2023. URL http://arxiv.org/abs/2303.14080. arXiv:2303.14080 [cs].

Hayler, A., Huang, X., Ceylan, , Bronstein, M., and Finkelshtein, B. Bringing Graphs to the Table: Zero-shot Node Classification via Tabular Foundation Models, 2025. URL https://arxiv.org/abs/2509.07143. Version Number: 2.

He, X., Zhao, K., and Chu, X. AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, January 2021. ISSN 0950-7051. doi: 10.1016/j.knosys.2020.106622. URL https://www.sciencedirect.com/science/article/pii/S0950705120307516.

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., and Sontag, D. TabLLM: Few-shot Classification of Tabular Data with Large Language Models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581. PMLR, April 2023. URL https://proceedings.mlr.press/v206/hegselmann23a.html. ISSN: 2640-3498.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. The Eleventh International Conference on Learning Representations, September 2022. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, January 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08328-6. URL https://www.nature.com/articles/s41586-024-08328-6. Publisher: Nature Publishing Group.

Hoo, S. B., Müller, S., Salinas, D., and Hutter, F. The tabular foundation model tabpfn outperforms specialized time series forecasting models based on simple features. In *NeurIPS workshop on time series in the age of large models*, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. October 2021. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hu, W., Yuan, Y., Zhang, Z., Nitta, A., Cao, K., Kocijan, V., Sunil, J., Leskovec, J., and Fey, M. PyTorch Frame: A Modular Framework for Multi-Modal Tabular Learning. October 2024. URL https://openreview.net/forum?id=2ZHKA9xo8V#discussion.

Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., and Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3 (1):136, October 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00341-z. URL https://www.nature.com/articles/s41746-020-00341-z.

Jiang, J.-P., Ye, H.-J., Wang, L., Yang, Y., Jiang, Y., and Zhan, D.-C. Tabular Insights, Visual Impacts: Transferring Expertise from Tables to Images. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 21988–22009. PMLR, July 2024. URL https://proceedings.mlr.press/v235/jiang24h.html.

Jiang, J.-P., Liu, S.-Y., Cai, H.-R., Zhou, Q.-L., and Ye, H.-J. Representation Learning for Tabular Data: A Comprehensive Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2026. ISSN 1939-3539. doi: 10.1109/TPAMI.2026.3657217. URL https://ieeexplore.ieee.org/abstract/document/11369258.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/644

9f44a102fde848669bdd9eb6b76fa-Abstract.html.

Khattab, O. and Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pp. 39–48, New York, NY, USA, July 2020. Association for Computing Machinery. ISBN 978-1-4503-8016-4. doi: 10.1145/3397271.3401075. URL https://dl.acm.org/doi/10.1145/3397271.3401075.

Kim, M. J., Grinsztajn, L., and Varoquaux, G. CARTE: pre-training and transfer for tabular learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 23843–23866, Vienna, Austria, July 2024. JMLR.org.

Kim, M. J., Lefebvre, F., Brison, G., Perez-Lebel, A., and Varoquaux, G. Table Foundation Models: on knowledge pre-training for tabular learning, May 2025a. URL http://arxiv.org/abs/2505.14415. arXiv:2505.14415 [cs].

Kim, W., Song, C., and Kim, H. MultiModalPFN: Extending Prior-Data Fitted Networks for Multimodal Tabular Learning. October 2025b. URL https://openreview.net/forum?id=pSyuFl8mau.

Koshorek, O., Granot, N., Alloni, A., Admati, S., Hendel, R., Weiss, I., Arazi, A., Cohen, S.-N., and Belinkov, Y. Structured RAG for Answering Aggregative Questions, November 2025. URL http://arxiv.org/abs/2511.08505. arXiv:2511.08505 [cs].

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.

Li, W., Tang, R., Li, C., Zhang, C., Vulić, I., and Søgaard, A. Lost in Embeddings: Information Loss in Vision–Language Models. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22676–22693, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1235. URL https://aclanthology.org/2025.findings-emnlp.1235/.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Liu, Y., Zhang, Y., Ghosh, D., Schmidt, L., and Yeung-Levy, S. Data or Language Supervision: What Makes CLIP Better than DINO? In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1868–1874, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.98. URL https://aclanthology.org/2025.findings-emnlp.98/.

Lu, J., Qian, Y., Zhao, S., Xi, Y., and Yang, C. MuG: A Multimodal Classification Benchmark on Game Data with Tabular, Textual, and Visual Fields. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5332–5346, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.354. URL https://aclanthology.org/2023.findings-emnlp.354/.

Luo, A., Du, J., Tian, F., Xian, X., Specht, R., Wang, G., Bi, X., Fleming, C., Srinivasa, J., Kundu, A., Hong, M., and Ding, J. Can Agentic AI Match the Performance of Human Data Scientists?, December 2025a. URL http://arxiv.org/abs/2512.20959. arXiv:2512.20959 [cs].

Luo, J., Yuan, Y., and Xu, S. TIME: TabPFN-Integrated Multimodal Engine for Robust Tabular-Image Learning, June 2025b. URL http://arxiv.org/abs/2506.00813. arXiv:2506.00813 [cs].

Ma, J., Thomas, V., Hosseinzadeh, R., Kamkari, H., Labach, A., Cresswell, J. C., Golestan, K., Yu, G., Caterini, A. L., and Volkovs, M. TabDPT: Scaling Tabular Foundation Models on Real Data, July 2025. URL http://arxiv.org/abs/2410.18164. arXiv:2410.18164 [cs].

Malaviya, C., Shaw, P., Chang, M.-W., Lee, K., and Toutanova, K. QUEST: A Retrieval Dataset of Entity-Seeking Queries with Implicit Set Operations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14032–14047, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.784. URL https://aclanthology.org/2023.acl-long.784.

McElfresh, D., Khandagale, S., Valverde, J., Prasad C, V., Ramakrishnan, G., Goldblum, M., and White, C. When Do Neural Nets Outperform Boosted Trees on Tabular Data? *Advances in Neural Information Processing Systems*, 36:76336–76369, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f06d5ebd4ff40

b40dd97e30cee632123-Abstract-Datasets_and_Benchmarks.html.

Meghawat, M., Yadav, S., Mahata, D., Yin, Y., Shah, R. R., and Zimmermann, R. A multimodal approach to predict social media popularity. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 190–195. IEEE, 2018.

Mráz, M., Das, B., Gupta, A., Purucker, L., and Hutter, F. Towards Benchmarking Foundation Models for Tabular Data With Text. June 2025. URL https://openreview.net/forum?id=yrmoQG9NAV.

Müller, S., Hollmann, N., Arango, S. P., Grabocka, J., and Hutter, F. Transformers Can Do Bayesian Inference. October 2021. URL https://openreview.net/forum?id=KSugKcbNf9.

Ophir, Y., Tikochinski, R., Asterhan, C. S., Sisso, I., and Reichart, R. Deep neural networks detect suicide risk from textual facebook posts. *Scientific reports*, 10(1): 16685, 2020.

Pantazopoulos, G., Suglia, A., Lemon, O., and Eshghi, A. Lost in Space: Probing Fine-grained Spatial Understanding in Vision and Language Resamplers. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 540–549, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.45. URL https://aclanthology.org/2024.naacl-short.45/.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.

Pu, Y., He, Z., Jiang, Y., Qiu, T., Wu, H., Sun, Q., Zhuo, C., and Yu, B. Customized Retrieval Augmented Generation and Benchmarking for EDA Tool Documentation QA. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 44(12):4615–4628, December 2025. ISSN 1937-4151. doi: 10.1109/TCAD.2025.3568776. URL https://ieeexplore.ieee.org/abstract/document/10994463.

Qu, J., Holzmüller, D., Varoquaux, G., and Morvan, M. L. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data, February 2025. URL http://arxiv.org/abs/2502.05564. arXiv:2502.05564 [cs].

Qu, J., Holzmüller, D., Varoquaux, G., and Morvan, M. L. TabICLv2: A better, faster, scalable, and open tabular foundation model, February 2026. URL http://arxiv.org/abs/2602.11139. arXiv:2602.11139 [cs].

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, July 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Robertson, J., Reuter, A., Guo, S., Hollmann, N., Hutter, F., and Schölkopf, B. Do-PFN: In-Context Learning for Causal Effect Estimation, 2025. URL https://arxiv.org/abs/2506.06039. Version Number: 3.

Schindler, G., Schambach, M., Medek, M., and Thelin, S. TabGemma: Text-Based Tabular ICL via LLM using Continued Pretraining and Retrieval. November 2025. URL https://openreview.net/forum?id=cv2uilalQx.

Shapira, E., Madmon, O., Reichart, R., and Tennenholtz, M. Can llms replace economic choice prediction labs? the case of language-based persuasion games. *arXiv preprint arXiv:2401.17435*, 2024.

Shi, X., Mueller, J., Erickson, N., Li, M., and Smola, A. Benchmarking Multimodal AutoML for Tabular Data with Text Fields. August 2021. URL https://openreview.net/forum?id=Q0zOIaec8HF.

Shwartz-Ziv, R. and Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, May 2022. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.11.011. URL https://www.sciencedirect.com/science/article/pii/S1566253521002360.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., and Bojanowski, P. DINOv3, August 2025. URL http://arxiv.org/abs/2508.10104. arXiv:2508.10104 [cs].

Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A. J., Ananthram, A., Nathan, A., Luo, A., Helyar, A., Madry, A., Efremov, A., Spyra, A., Baker-Whitcomb, A., Beutel, A., Karpenko, A., Makelov, A., et al. OpenAI GPT-5 System Card, December 2025. URL https://arxiv.org/abs/2601.03267v1.

Spathis, D. and Kawsar, F. The first step is the hardest: pitfalls of representing and tokenizing temporal data for large language models. *Journal of the American Medical Informatics Association*, 31(9):2151–2158, September 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae090. URL https://doi.org/10.1093/jamia/ocae090.

Spinaci, M., Polewczyk, M., Schambach, M., and Thelin, S. ConTextTab: A Semantics-Aware Tabular In-Context Learner, June 2025. URL http://arxiv.org/abs/2506.10707. arXiv:2506.10707 [cs].

Sukel, M., Rudinac, S., and Worring, M. Multimodal temporal fusion transformers are good product demand forecasters. *IEEE MultiMedia*, 31(2):48–60, 2024.

Tang, Y. and Yang, Y. Do We Need Domain-Specific Embedding Models? An Empirical Investigation. October 2024. URL https://openreview.net/forum?id=powufeT93G.

Tang, Z., Fang, H., Zhou, S., Yang, T., Zhong, Z., Hu, C., Kirchhoff, K., and Karypis, G. AutoGluon-Multimodal (AutoMM): Supercharging Multimodal AutoML with Foundation Models. AutoML Conference 2024 (ABCD Track), April 2024a. URL https://openreview.net/forum?id=irStSm9waW.

Tang, Z., Zhong, Z., He, T., and Friedland, G. Bag of Tricks for Multimodal AutoML with Image, Text, and Tabular Data, December 2024b. URL http://arxiv.org/abs/2412.16243. arXiv:2412.16243 [cs].

Thawani, A., Pujara, J., Ilievski, F., and Szekely, P. Representing Numbers in NLP: a Survey and a Vision. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–656, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.53. URL https://aclanthology.org/2021.naacl-main.53/.

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., and Xie, S. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. pp. 9568–9578, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tong_Eyes_Wide_Shut_Exploring_the_Visual_Shortcomings_of_Multimodal_LLMs_CVPR_2024_paper.html.

Van Breugel, B. and Van Der Schaar, M. Position: why tabular foundation models should be a research priority. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 48976–48993, Vienna, Austria, July 2024. JMLR.org.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, F., Li, Y., and Xiao, H. jina-reranker-v3: Last but Not Late Interaction for Listwise Document Reranking, October 2025. URL http://arxiv.org/abs/2509.25085. arXiv:2509.25085 [cs].

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. Text Embeddings by Weakly-Supervised Contrastive Pre-training, February 2024. URL http://arxiv.org/abs/2212.03533. arXiv:2212.03533 [cs].

Weller, O., Boratko, M., Naim, I., and Lee, J. On the Theoretical Limitations of Embedding-Based Retrieval. October 2025. URL https://openreview.net/forum?id=k9CzIvzfaA.

Wu, J., Gan, W., Chen, Z., Wan, S., and Yu, P. S. Multimodal Large Language Models: A Survey, November 2023. URL http://arxiv.org/abs/2311.13165. arXiv:2311.13165 [cs].

Yan, J., Zheng, B., Xu, H., Zhu, Y., Chen, D., Sun, J., Wu, J., and Chen, J. Making Pre-trained Language Models Great on Tabular Prediction. The Twelfth International Conference on Learning Representations, October 2023. URL https://openreview.net/forum?id=anzIzGZuLi.

Ye, H.-J., Liu, S.-Y., Cai, H.-R., Zhou, Q.-L., and Zhan, D.-C. A Closer Look at Deep Learning Methods on Tabular Datasets, November 2025. URL http://arxiv.org/abs/2407.00956. arXiv:2407.00956 [cs].

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, December 2024. ISSN 2095-5138. doi: 10.1093/nsr/nwae403. URL https://doi.org/10.1093/nsr/nwae403.

Zhang, X., Maddix, D. C., Yin, J., Erickson, N., Ansari, A. F., Han, B., Zhang, S., Akoglu, L., Faloutsos, C., Mahoney, M. W., Hu, C., Rangwala, H., Karypis, G., and Wang, B. Mitra: Mixed Synthetic Priors for Enhancing Tabular Foundation Models, October 2025a. URL http://arxiv.org/abs/2510.21204. arXiv:2510.21204 [cs].

Zhang, X., Ren, G., Yu, H., Yuan, H., Wang, H., Li, J., Wu, J., Mo, L., Mao, L., Hao, M., Dai, N., Xu, R., Li, S., Zhang, T., He, Y., Wang, Y., Zhang, Y., Xu, Z., Li, D., Gao, F., Zou, H., Liu, J., Liu, J., Xu, J., Cheng, K., Li, K., Zhou, L., Li, Q., Fan, S., Lin, X., Han, X., Li, X., Lu, Y., Xue, Y., Jiang, Y., Wang, Z., Wang, Z., and Cui, P. LimiX: Unleashing Structured-Data Modeling Capability for Generalist Intelligence, November 2025b. URL http://arxiv.org/abs/2509.03505. arXiv:2509.03505 [cs].

A. Extended Introduction

Tabular Foundation Models (TFMs) (Van Breugel & Van Der Schaar, 2024; Hollmann et al., 2022; 2025; Qu et al., 2025; Grinsztajn et al., 2026; Qu et al., 2026) have recently emerged as the state of the art (SOTA) for supervised tabular learning (Erickson et al., 2025; Ye et al., 2025). They have surpassed gradient-boosted decision trees (GBDTs) (Breiman, 2001; Chen & Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018), which have historically been the leading approach (Shwartz-Ziv & Armon, 2022; Grinsztajn et al., 2022; McElfresh et al., 2023). Recently, these versatile learners have been extended to causal inference (Robertson et al., 2025), graph learning (Hayler et al., 2025), and time-series (Hoo et al., 2024). However, the best-performing TFMs (Grinsztajn et al., 2026; Qu et al., 2026) are trained exclusively on structured numerical data, making them fundamentally unimodal: unstructured inputs must be preprocessed via external embedding models (Wang et al., 2024; Siméoni et al., 2025), with no unified support for modalities such as text and image.

Yet, in many high-impact domains, tabular problems are multimodal: e-commerce listings (Das et al., 2024; Sukel et al., 2024; Shapira et al., 2024), social media feeds (Meghawat et al., 2018; Ophir et al., 2020; Badian et al.), and medical health records (Huang et al., 2020; Cui et al., 2023; Duenias et al., 2025; Fu et al., 2025) combine image and text with numerical features. While early work has begun extending TFMs to integrate text (Arazi et al., 2025; Spinaci et al., 2025), these extensions often compromise the model's core tabular performance, and inherent support for visual modalities remains entirely absent. One might turn to Large Language and Vision-Language Models (LLMs/VLMs), which natively process unstructured inputs, but they are not suited for the inductive biases of tabular data; specifically, they are unoptimized for the relational structure (Fang et al., 2024) and are suboptimal for numerical features (Van Breugel & Van Der Schaar, 2024). Addressing these limitations requires architectures that combine the numerical precision of TFMs while maintaining the rich input handling of multimodal foundation models. However, evaluating such a unified approach is difficult because the diverse nature of tasks within Multimodal Tabular Learning (MMTL) (Jiang et al., 2026; Kim et al., 2025b) is not yet fully characterized; existing benchmarks (Shi et al., 2021; Lu et al., 2023; Kim et al., 2024; Tang et al., 2024b; Mráz et al., 2025) primarily highlight the coexistence of modalities, unintentionally grouping together problems that require fundamentally different modeling solutions.

To characterize these problems, we observe that tabular models require inputs to be represented as feature columns, so high-dimensional images and texts must be compressed into compact representations. Consequently, embeddings act as lossy summaries, as they capture only a fraction of the raw input's information by design (Weller et al., 2025). In order to generalize well, pretrained embedding models are optimized for broad semantic content, such as distinguishing an X-ray from a mammogram, at the expense of fine-grained details like precise size estimations or localized anomalies (Pantazopoulos et al., 2024; Li et al., 2025). While this compression is effective for global semantic mapping, it fails to preserve the specialized signals required for fine-grained MMTL tasks. For example, the optimal representation of a chest X-ray differs depending on whether the tabular task is to diagnose pneumonia or a rib fracture, and whether the patient is a young athlete or an elderly smoker. We thus advocate for the need for Target-Aware Representations (TAR): embeddings that are tuned to the target and, ideally, to the other modalities.

Consider, for example, the task of pneumonia detection from a patient record combining age and smoking status with chest X-ray images. We argue that to study MMTL, a dataset should satisfy two properties: (1) *Joint Signal*, where each modality provides complementary information that contributes to the overall predictive performance, and (2) *Task-awareness*, where task-agnostic representations fail to capture the details required for a given objective. In our example, both the X-ray and the clinical profile offer unique, complementary information, and steering the image embedding to detect subtle signs of inflammation in the lungs should improve diagnostic accuracy.

To translate these theoretical properties into a measurable test, we develop an algorithmic pipeline that quantifies whether a dataset complies with the aforementioned requirements. This approach approximates these properties by evaluating each task across a broad suite of tabular learners, ranging from light GBDTs to SOTA TFMs. To evaluate for *Joint Signal*, we demand a performance drop when either modality is removed, verifying that each input strengthens the predictive power. For *Task-awareness*, we finetune the encoder's last 3 layers with LoRA (Hu et al., 2021) on the prediction target as a preprocessing step, and we expect these representations to outperform frozen ones when passed to tabular models. Crucially, our experiments confirm that target-aware representations outperform frozen embeddings across established MMTL benchmarks; however, we find that the magnitude of these gains is highly dataset-dependent, suggesting they represent distinct classes of MMTL tasks.

Building on this framework, we introduce **MulTaBench**, a benchmark of 40 datasets balanced between image-tabular and text-tabular tasks, as well as classification and regression objectives. To ensure a comprehensive evaluation, the benchmark

incorporates a wide range of sample sizes and feature counts, while spanning a diverse set of domains to capture the heterogeneity of real-world multimodal tabular data. MulTaBench represents the largest image-tabular benchmarking effort to date, and the first MMTL benchmark to explicitly prioritize datasets requiring task-aware representations. Demonstrating the robustness of our curation criteria, we show that the gains from target-aware tuning generalize consistently across a diverse suite of independent tabular learners, encoder scales, and embedding dimensions. These findings suggest that designing novel architectures which contextualize the representations of unstructured modalities can push the boundaries of MMTL, and we believe that MulTaBench would be instrumental for developing true Multimodal TFMs.

B. Related Work

Tabular Foundation Models. The landscape of tabular learning shifted with Prior-data Fitted Networks (PFNs) (Müller et al., 2021), which pretrain transformers over synthetic tabular datasets with in-context learning (ICL) (Brown et al., 2020). The TabPFN family (Hollmann et al., 2022; 2025; Grinsztajn et al., 2026; Garg et al., 2025) pioneered this direction. Multiple subsequent works (Qu et al., 2025; 2026; Ma et al., 2025; Zhang et al., 2025a; Spinaci et al., 2025; Zhang et al., 2025b; Bouadi et al., 2025) advanced the paradigm with improvements spanning synthetic data diversity, real-world data pretraining, and architectural scalability. Among these, ConTextTab (Spinaci et al., 2025) is the only PFN to incorporate textual fields, yet it does not process raw strings; instead, it relies on external, frozen text embeddings as static inputs, decoupling the representation from the tabular learning objective. In addition, several non-PFN approaches (Yan et al., 2023; Kim et al., 2024; 2025a) also incorporate semantic awareness, but likewise treat text representations as frozen. TabSTAR (Arazi et al., 2025) represents a fundamental shift: rather than processing fixed representations, it jointly trains both the textual and tabular encoders, successfully demonstrating that TAR are essential for MMTL. However, it lacks support for images and its non-ICL architecture compromises its numerical performance.

LLMs and VLMs. Recent years have seen the rise of LLMs and their evolution into VLMs (Wu et al., 2023; Yin et al., 2024; Caffagni et al., 2024). These powerful models (Singh et al., 2025; Comanici et al., 2025) typically employ a unified transformer architecture (Vaswani et al., 2017) to process interleaved modalities within a single sequence, offering a path to integrate tabular data with text and image; however, research has primarily focused on text-tabular tasks (Fang et al., 2024). TabLLM (Hegselmann et al., 2023) explored different strategies to serialize the tabular data into natural language, and TabuLa-8B (Gardner et al., 2024) and TabGemma (Schindler et al., 2025) combined continued pretraining of LLMs on tabular corpora (Eggert et al., 2023) with architectural modifications, achieving strong few-shot performance. Nevertheless, the autoregressive nature of LLMs is misaligned with the structure of tabular data, and their tokenization process damages numerical precision (Thawani et al., 2021; Spathis & Kawsar, 2024). Furthermore, their massive scale introduces prohibitive costs for high-throughput inference, while their extensive pretraining risks memorizing evaluation data (Bordt et al., 2024; Gorla & Puduppully, 2026). Consequently, generative architectures remain largely impractical for discriminative MMTL.

Joint Multimodal Tabular Learning Architectures. Despite various architectural proposals (Hager et al., 2023; Jiang et al., 2024; Ebrahimi et al., 2024; Hu et al., 2024; Du et al., 2025), the field still lacks a true multimodal foundation model for tabular data with text and images. AutoML (He et al., 2021) frameworks (Shi et al., 2021; Tang et al., 2024a;b), led by AutoGluon-Multimodal (Tang et al., 2024a), demonstrated the benefit of joint modeling by combining tabular, text and image encoders. However, their reliance on a non-ICL transformer (Gorishniy et al., 2021) as the tabular backbone limits their tabular capacities. Similarly, TabSTAR (Arazi et al., 2025) introduced a jointly pretrained text-tabular architecture and achieved strong performance on text-tabular classification tasks, but it struggled with regression tasks and with unimodal tabular benchmarks (Erickson et al., 2025). Recent attempts have built on stronger tabular foundations, by expanding the PFN paradigm with multimodal fusion strategies. TIME (Luo et al., 2025b) proposed a late-fusion approach in an image-tabular setup, but missed cross-modal interactions and achieved mixed results when employing finetuning. MultiModalPFN (Kim et al., 2025b) fused TabPFN with visual and textual backbones, but assumed frozen multimodal embeddings. To conclude, no existing model has successfully maintained SOTA performance on tabular tasks while learning TAR for text and images.

Text-Tabular Benchmarks. Existing text-tabular benchmarks differ significantly in their curation philosophy and dataset scale. The Multimodal AutoML Benchmark (Shi et al., 2021) introduced 18 datasets with deliberate diversity in task type and predictive signal. Grinsztajn et al. (2023) filtered 14 datasets from a bigger pool, where the text features provided a significant gain over a numerical-only baseline. TextTabBench (Mráz et al., 2025) curated 13 text-tabular datasets, focusing on longer text fields while ensuring both the text modality and numerical features contribute to the prediction. CARTE (Kim et al., 2024) collected 51 datasets, mainly featuring short strings and high-cardinality categories, typically present

in knowledge graphs. While these efforts were instrumental in advancing research on tabular data with strings, none of them were deliberately designed to isolate tasks where static representations fail to capture the necessary predictive signal. Importantly, as we show in § 3, most of the datasets included in the aforementioned benchmarks do not pass our curation pipeline. Consequently, potential performance gains that native Multimodal TFMs are designed to deliver might be overlooked. For example, ConTextTab set the SOTA for the CARTE benchmark (Spinaci et al., 2025), but struggles on MulTaBench (see § 4).

Image-Tabular Benchmarks. The availability of image-tabular benchmarks is highly limited. MuG (Lu et al., 2023) introduced 4 data sources from the gaming domain combining tabular data with text and image, but offering limited domain diversity. Similarly, Tang et al. (2024b) curated 11 tabular datasets with images, but without quantifying the image signal's necessity. As detailed in § 3, these datasets often fail our curation pipeline and suffer from additional quality issues. The lack of large accessible benchmarks led recent work such as TIME (Luo et al., 2025b) and MultimodalTabPFN (Kim et al., 2025b), to rely on a self-selected group of datasets, limiting the generalizability of their findings. We address this gap by doubling the benchmark size and assuring that the image representations are central for MMTL.

Limits of Frozen Representations. Pretrained representations are optimized for general-purpose objectives and often fail to capture the fine-grained, task-specific details necessary for downstream performance (Tong et al., 2024; Liu et al., 2025; Gisserot-Boukhlef et al., 2025; Cao et al., 2026). Weller et al. (2025) provide a theoretical basis for this limitation, demonstrating how RAG systems (Lewis et al.) that rely on static embeddings can fail on even seemingly simple cases. To overcome this problem, alternative approaches (Khattab & Zaharia, 2020; Malaviya et al., 2023; Fan et al., 2024; Tang & Yang, 2024; Edge et al., 2025; Wang et al., 2025; Pu et al., 2025; Koshorek et al., 2025) enabled the contextualization of document representations in the presence of the query. Similar limitations were also illustrated in VQA (Antol et al., 2015), where encoding images independently of the question leads to information loss, as the query determines which image regions are predictive (Ganz et al., 2024; Li et al., 2025). To overcome these limitations, VLMs have evolved toward deep multimodal alignment (Radford et al., 2021; Alayrac et al., 2022; Liu et al., 2023), and we argue that MMTL should undergo a similar evolution, moving away from decoupled preprocessing and frozen embeddings in favor of a joint learning approach.

C. Curation Pipeline

C.1. Target-Aware Representations

Target-Aware Representations are produced by finetuning the top 3 transformer layers of the encoder using LoRA (Hu et al., 2021), with a single linear head mapping the encoder output (384-dim) to the number of output classes. Finetuning is performed as a preprocessing step, independently of the structured features and the downstream tabular learner. The encoder is adapted on the training split only, using a stratified 90/10 train/validation split to select the best checkpoint. Importantly, there is no data leakage, as the test set is never used for this step, just like any other preprocessing.

Hyperparameters. Both DINO-v3-small[1] and e5-small-v2[2] share the same LoRA configuration: $r = 16$, $\alpha = 32$, dropout 0.1. Training uses AdamW with learning rate 10^{-4} for e5 and 0.001 for DINO, with a batch size of 256, and weight decay 0.01. For DINO, we train to up to 100 epochs. As many datasets have multiple text features, we reduce this number for e5 to 50. We apply early stopping after 3 epochs of no improvement on the validation loss. All hyperparameters are fixed across datasets; no per-dataset tuning is performed. Reported gains are therefore conservative lower bounds on what task-specific adaptation could achieve.

Regression. For regression targets, the continuous label is discretized into 20 equal-frequency bins and the adaptation objective is cross-entropy over these bins. We find this technique to be more stable than direct regression finetuning, as it is much less sensitive to outliers. However, it's plausible that this decision could be optimized much further.

Text. While MulTaBench image datasets have a single image feature, text-tabular datasets often have more than one text field, which we defined as string features that have at least 100 distinct values. For efficiency, a single e5 model is finetuned jointly across all text columns: each row-col pair generates one training example in the format "$col\_name : col\_val$", paired

[1] https://huggingface.co/facebook/dinov3-vits16-pretrain-lvd1689m
[2] https://huggingface.co/intfloat/e5-small-v2

with the row's target label. This allows the model to learn a shared representation across all text features simultaneously. This decision might harm representations, especially as feature size grows, but finetuning a dedicated embedding model for each feature would have been computationally infeasible.

C.2. Curation Experimental Setup

Each candidate dataset is evaluated by 5 tabular learners: LightGBM, CatBoost, TabM, TabPFNv2, and TabPFN-2.5, over five random seeds under the 4 conditions defined in §2.2. Training is capped at 10,000 examples per fold, and the metric is AUC for classification and R^2 for regression tasks.

We run models using default configurations. For LightGBM, we use its default implementation[3]. For CatBoost, we follow previous work (Gorishniy et al., 2021; Arazi et al., 2025) and set $early\_stopping\_rounds = 50, od\_pval = 0.001, iterations = 2000$. For TabM, we use its *pytabkit*[4] implementation with default parameters. For TabPFNv2 and TabPFN-2.5, we use their default implementation.[5]

Table 1. Experimental Conditions. Breakdown by feature composition and representation strategy.

Condition	Structured	Unstructured	Target-Aware (TAR)
Unimodal Structured	✓	×	–
Unimodal Unstructured	×	✓	×
Joint Frozen	✓	✓	×
Joint TAR	✓	✓	✓

C.3. Formal Acceptance Criteria

Let $\mathcal{D}$ be a candidate dataset and $\mathcal{M}$ be a pool of 5 curation tabular learners. For a given learner $m \in \mathcal{M}$, let $S_m(\text{Condition})$ denote its average predictive performance (AUC or R^2) under a given condition.

Joint Signal. We define the *Joint* gain as the improvement of the joint model over the strongest unimodal baseline:

$$\Delta_{\text{Joint}}(m) = S_m(\text{Joint Frozen}) - \max\left(S_m(\text{UnimodalStructured}),\ S_m(\text{UnimodalUnstructured})\right)$$

Task-awareness. We define the *Awareness* gain as the improvement of Joint TAR over Joint Frozen:

$$\Delta_{\text{Awareness}}(m) = S_m(\text{Joint TAR}) - S_m(\text{Joint Frozen})$$

Selection rule. To ensure that the observed improvements are robust and exceed a minimum significance margin, we introduce a threshold parameter $\delta \geq 0$ and a consensus fraction $\rho \in (0.5, 1]$. A dataset $\mathcal{D}$ is accepted if and only if both gains exceed δ for a majority of the learners:

$$\text{Accept}(\mathcal{D}) \iff \left|\{m \in \mathcal{M} \ :\ \Delta_{\text{Joint}}(m) > \delta \ \wedge\ \Delta_{\text{Awareness}}(m) > \delta\}\right| \geq \rho \cdot |\mathcal{M}|$$

The two conditions are evaluated jointly per learner: a model counts toward the threshold only if both gains are above the threshold. In our case, we set $\delta = 0.001$ and $\rho = 3/5$. We note that since we use a binary decision threshold, some datasets can marginally cross it while others can generate consensus. Demanding stricter thresholds could enhance the robustness of the selected datasets.

[3] https://pypi.org/project/lightgbm/
[4] https://github.com/dholzmueller/pytabkit
[5] https://github.com/PriorLabs/TabPFN

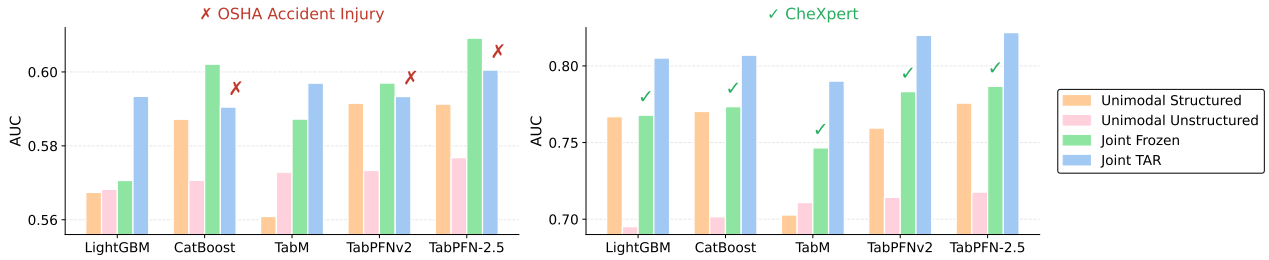

Figure 4. Curation protocol over candidate datasets. Mean AUC per model and condition. The *OSHA Accident Injury* dataset is rejected as *TAR* fails to consistently improve over *Joint Frozen*.

D. MulTaBench Datasets

In this section we present MulTaBench datasets. Table 2 provides their high-level statistics, including the number of rows and feature type breakdown. The rest of the section provides a concise per-dataset high-level description; their exact preprocessing logic can be found in our released code.

D.1. Image-Tabular Dataset Descriptions

CBIS-DDSM. Cropped mammography mass regions from the Curated Breast Imaging Subset of DDSM, with 1,696 crops. The 4-class target is BI-RADS breast density (categories 1–4). Structured features describe lesion morphology, such as laterality, imaging view (MLO or CC), mass shape, mass margins, BI-RADS assessment score, pathology (malignant/benign), and subtlety rating.

Celeb Attractiveness. Celebrity face images from the CelebA dataset, sampled to 99,999 images[6] from the full 202,599. The binary target is a crowd-annotated attractiveness label. Each row pairs the face image with 39 binary facial-attribute features, such as *Smiling*, *Wearing_Lipstick* or, *Bald*, making the image a complement to an already rich structured signal.

CheXpert. Chest X-ray images from the Stanford CheXpert dataset, with 46,437 frontal and lateral views. The 3-class target predicts Cardiomegaly label (positive, negative, or uncertain). Structured features include patient sex, age, and 14 co-occurring pathology labels, many of which are sparsely observed (over 85% missing for several conditions), reflecting natural label uncertainty in radiology reports.

CS:GO Skins. Weapon skin images and metadata from the Counter-Strike: Global Offensive marketplace, with 956 cosmetic items. The 10-class target discretizes market price into decile quantile bins. Structured features include skin quality (rarity tier), weapon category, and availability; a free-text skin name column provides additional descriptive signal about the skin's visual design.

Flower Bouquets. Flower bouquet photographs paired with sales metadata from a Russian online florist, comprising 600 listings. The 5-class target is a customer satisfaction rating (1–5). Features include a free-text bouquet description, average comment-based rating, and price.

Glaucoma SMDG. Retinal fundus photographs from the SMDG multi-source glaucoma benchmark, with 12,449 images. The 3-class target encodes glaucoma diagnosis (positive, negative, or uncertain).Clinical metadata including patient age, sex, laterality, and intraocular pressure are available as structured features, though heavily sparse (over 99% missing for several fields), reflecting real-world incompleteness in ophthalmic records.

Hateful Meme. Multimodal memes from the Facebook Hateful Memes Challenge, comprising 10,000 image-text pairs. The binary target labels each meme as hateful or not. To make it a tabular task, we pre-embedded the text field into 20 continuous variables, to capture part (but not all) of the text signal. The structured columns should thus be treated as numeric features rather than raw text, with the meme image providing complementary visual context.

[6]This was originally intended to be 100,000, but we eventually dropped an observation with a corrupted image.

Table 2. All 40 MulTaBench Datasets Properties. *Task*: Classification (CLS) or Regression (REG). *Classes*: number of target classes (for CLS). *N*: total examples. *Struct.*: numerical + categorical features. *Text*: free-text features. *Img.*: image features.

Dataset	Task	Classes	N	Struct.	Text	Img.
Image-Tabular (20 datasets)						
CBIS-DDSM	CLS	4	1,696	8	0	1
Celeb Attractiveness	CLS	2	99,999	39	0	1
CheXpert	CLS	3	46,437	17	0	1
CS:GO Skins	CLS	10	956	3	1	1
Flower Bouquets	CLS	5	600	3	1	1
Glaucoma SMDG	CLS	3	12,449	8	0	1
Hateful Meme	CLS	2	10,000	20	0	1
HubMAP HPA	CLS	10	12,581	3	1	1
Justin Instagram	CLS	5	10,319	6	0	1
Mammography CMMD	CLS	2	5,202	4	0	1
PetFinder	CLS	8	14,652	17	4	1
Zooscan Plankton	CLS	10	100,000	28	0	1
Amazon Bestseller	REG	–	3,488	4	0	1
Amazon Packages	REG	–	46,398	1	1	1
H&M Fashion	REG	–	104,072	9	4	1
Khaadi Clothes	REG	–	400	2	1	1
Letterboxd Movies	REG	–	12,564	23	3	1
Mango Mass	REG	–	546	2	0	1
MkPhoto Bots	REG	–	13,748	8	0	1
Painting Price	REG	–	12,369	245	2	1
Text-Tabular (20 datasets)						
Data Scientist Salary	CLS	6	15,841	1	5	0
Fake Job Postings	CLS	2	12,725	2	3	0
Jigsaw Toxicity	CLS	2	100,000	29	2	0
Kickstarter	CLS	2	86,502	4	5	0
Michelin Guide	CLS	5	18,843	5	6	0
Product Sentiment	CLS	4	5,091	1	1	0
Spotify Genres	CLS	114	114,000	15	3	0
US Accidents	CLS	4	100,001	35	9	0
Wine Review	CLS	30	84,123	3	2	0
Women's Clothing	CLS	5	18,788	8	2	0
Baby Products	REG	–	5,085	8	4	0
Book Price	REG	–	4,989	3	5	0
Book Readability	REG	–	4,724	24	6	0
Mercari Marketplace	REG	–	100,000	3	6	0
Montgomery Salaries	REG	–	9,228	7	4	0
Rotten Tomatoes	REG	–	7,158	2	13	0
SciMagojr Impact	REG	–	31,136	12	10	0
Vancouver Salaries	REG	–	44,574	3	2	0
Video Games Sales	REG	–	16,598	3	2	0
Zomato Restaurants	REG	–	41,665	8	7	0

HubMAP HPA. Histology tissue tile images from the HuBMAP-HPA organ segmentation competition, with 12,581 tiles. The 10-class target discretizes donor age into decile quantile bins, asking whether tissue morphology encodes biological age. Structured features include organ type (kidney, prostate, large intestine, spleen, lung), donor sex, and tile coordinates; a run-length encoding column of the segmentation mask is present but largely absent (61% missing).

Justin Instagram. Instagram posts from five celebrities named Justin (Bieber, Trudeau, Timberlake, Long, Hartley), totaling 10,319 posts. The 5-class target identifies which Justin authored each post. Structured features are post-level metadata: number of hashtags, characters, words, emojis, and mentions, plus a binary video indicator.

Mammography CMMD. Mammography images from the Chinese Mammography Database, with 5,202 cropped lesion regions. The binary target distinguishes malignant from benign findings. Structured features include patient age, laterality (left/right), abnormality type (mass, calcification, or both), and the cropping method used (YOLO or contour detection).

PetFinder. Pet adoption listings from the Malaysian PetFinder platform, with 14,652 entries. The 8-class target discretizes listed pet age into octile bins, testing whether visual appearance and listing text jointly predict developmental stage. Features include species (cat/dog), breed, color, health status (vaccinated, dewormed, sterilized), adoption fee, state location, and a free-text listing description alongside the pet's photograph.

Zooscan Plankton. Underwater zooplankton specimens from the PELGAS Bay of Biscay survey, scanned with a ZooScan optical system, totaling 100,000 specimens. The 10-class target classifies copepod taxa (Calanoida, Oithonidae, Calanidae, Temoridae, and others). Structured features include 28 morphometric descriptors computed from the scan (circularity, skewness, fractal dimension, symmetry scores, area coverage, etc.) alongside sampling metadata such as geographic coordinates, depth, collection date, and mesh size.

Amazon Bestseller. Product listings from Amazon's bestseller rankings across all departments, with 3,488 items. The target is log-transformed product price. Structured features are the number of ratings, bestseller rank within department, star rating, and list page; the product thumbnail image provides visual cues about item type and packaging.

Amazon Packages. Warehouse bin images from Amazon's robotic fulfillment centers, with 46,398 bins. The target is the total weight of the bin's contents in pounds. The sole structured feature is the expected item count; a free-text product description column names the item in each bin.

H&M Fashion. Clothing article metadata and thumbnail images from H&M's product catalog, with 104,072 articles. The target is the average age of purchasing customers, capturing whether visual style and descriptive text encode demographic appeal. Structured attributes include product type, color group, graphical appearance, garment group, and department; text features are the product name and a free-text detail description, making this a trimodal dataset.

Khaadi Clothes. Apparel listings from the Pakistani fashion brand Khaadi, with 400 products. The target is retail price in Pakistani rupees. Structured features are color and product category; a free-text description column specifies fabric type and construction.

Letterboxd Movies. Film metadata and poster images from the Letterboxd movie-tracking platform, with 12,564 films released between 2021 and 2024. The target is the average community rating. Features include 19 binary genre flags, release year, runtime, and text fields for movie tagline and theme descriptions alongside the official poster image.

Mango Mass. Mango fruit photographs from a variety classification study, with 546 individual fruits. The target is fruit mass in kilograms. The only structured features are color group (yellow or green) and quality grade (1, 2, or premium), making the image the dominant signal for weight prediction.

MkPhoto Bots. Social media photographs collected for image authenticity research, with 13,748 posts. The target is a continuous trust score reflecting the estimated probability that the post is genuine. Structured features include binary flags for GAN generation and deepfake manipulation, presence of a person, face count, a recognized-celebrity list, upload speed, and a noise-quality score.

Painting Price. Painting images and metadata from an online art marketplace, with 12,369 works. The target is sale price. Structured features include physical dimensions (width, length), material (canvas, paper, wood, etc.), and 243 binary style tags (e.g., *abstract*, *impressionism*, *surrealism*); a high-cardinality free-text styles column provides additional stylistic signal.

D.2. Text-Tabular Dataset Descriptions

Data Scientist Salary. Indian data science job postings, with 15,841 listings. The 6-class target is salary band in lakh rupees per annum (0–3, 3–6, 6–10, 10–15, 15–25, 25–50). Text features include the experience range, job description (22% missing), job designation, required key skills, and city location; a noisy job-type field (75% missing) contributes as a weak structured signal.

Fake Job Postings. Job listings annotated for authenticity, with 12,725 postings. The binary target flags fraudulent listings. Text features include the job title and full description; structured features capture required experience level, required education, and salary range (83% missing), testing whether deceptive intent is expressed in free-text beyond coarse metadata.

Jigsaw Toxicity. Online comments from the Civil Comments platform, collected for Jigsaw's toxicity detection task, sampled to 100,000 instances. The binary target labels each comment as toxic. Alongside the comment text, structured features include 24 identity-mention fraction scores (e.g., proportions of annotators who identified references to religion, race, or gender; 77.5% missing) and five community-reaction counts (funny, wow, sad, likes, disagree).

Kickstarter. Crowdfunding campaigns from Kickstarter, with 86,502 projects. The binary target indicates whether the funding goal was reached. Text features are the project name, description, and keyword slug; structured features include the funding goal amount, country, currency, and campaign deadline and creation timestamps.

Michelin Guide. Restaurant listings from the 2021 Michelin Guide, with 18,843 restaurants worldwide. The 5-class target is the Michelin award level: Selected Restaurants, Bib Gourmand, and 1–3 Stars. Text features include restaurant name, address, city/country location, cuisine type, facilities and services, and a detailed Michelin editorial description; structured features are geo-coordinates, price tier, and a Green Star sustainability flag.

Product Sentiment. Tweets about Apple, Google, and Twitter products posted during SXSW 2011, with 5,091 posts. The 4-class target is sentiment: Positive, Negative, No Sentiment, or Cannot Say. The sole text feature is the tweet content; a numeric product-type column (10 integer-encoded product categories) identifies the product being discussed.

Spotify Genres. Spotify track metadata covering 1,000 tracks per genre across 114 genres, totaling 114,000 tracks. The 114-class target is the track genre. Text features include artist name, album name, and track name; structured features are 15 Spotify audio descriptors (danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and others).

US Accidents. Traffic accident records from the contiguous United States, sampled to 100,001 incidents. The 4-class target is accident severity on a 1–4 scale. Text features include a free-text incident description and eight location and weather text columns (street name, city, county, state, ZIP code, nearest airport code, weather condition, wind direction); structured features cover GPS coordinates, weather measurements, timestamps, and 12 binary road-feature flags.

Wine Review. Professional wine tasting notes from Wine Enthusiast magazine, with 84,123 reviews. The 30-class target is the grape variety. Text features are the tasting note description and province of origin; structured features are the numeric rating (points), price (6.6% missing), and country, making grape identification from flavor language a natural benchmark for text-tabular models.

Women's Clothing. Customer reviews of women's clothing from an anonymous US e-commerce retailer, with 18,788 reviews. The 5-class target is the star rating (1–5). Text features are the review title and full review text; structured features include customer age, product and department metadata, a binary recommendation indicator, and positive feedback count.

Baby Products. Nursery and baby product listings from a US retail catalog, with 5,085 items. The target is retail price. Text features are the product title, free-form brand name, and descriptive fields for color, fabric, and material (all sparsely

populated at 50–99% missing); structured features include a discount flag, product category, and physical dimensions (weight, length, width, height).

Book Price. Books listed on an online marketplace, with 4,989 titles. The target is log-transformed price in USD. Text features include the book title, author name, edition details, full synopsis, genre tag, and broad book category; structured features are average star rating and number of ratings (16.7% missing).

Book Readability. Text excerpts from the CLEAR Corpus, with 4,724 passages from children's and educational literature. The target is the New Dale–Chall Readability Formula score, a standard measure of text difficulty. The key text feature is the excerpt itself; structured features comprise 24 linguistic and bibliographic attributes including publication year, sentence and paragraph count, Flesch–Kincaid grade level, ARI, SMOG, and CAREC readability metrics, MPAA content rating, and Bradley–Terry easiness score.

Mercari Marketplace. Secondhand item listings from the Mercari mobile marketplace, sampled to 100,000 listings. The target is log-transformed sale price. Text features are the item name, free-text item description, and a three-level hierarchical category label; structured features are item condition (1–5), brand name (42.5% missing), and a binary shipping-included flag.

Montgomery Salaries. Annual salary records of Montgomery County (Maryland, USA) government employees, with 9,228 employees. The target is current annual salary. Text features include department name, division, job title, and underfilled title (88.2% missing); structured features are gender, 2016 gross pay, 2016 overtime pay (31.6% missing), assignment type (full/part time), and hire date.

Rotten Tomatoes. Movie metadata aggregated from IMDb and Rotten Tomatoes, with 7,158 films. The target is an audience/critic composite rating. Text features include movie name, director, screenwriter, full cast list, language, country, filming locations, genre tags, and plot description; structured features are release year, runtime, and rating and review counts.

SciMagojr Impact. Academic journal and book series metadata from the SCImago Journal & Country Rank database, with 31,136 entries. The target is the journal's H-index. Text features include the journal title, publisher name, coverage period, subject categories, and broad subject areas; structured features are the SJR impact score, quartile ranking, annual and three-year document and citation counts, Overton policy citation index, and SDG alignment score.

Vancouver Salaries. Annual salary disclosures for City of Vancouver public employees, with 44,574 records spanning 2007–2024. The target is annual remuneration. Text features are job title and department name; structured features are fiscal year, employee name, and declared expenses (5.4% missing).

Video Games Sales. Video game sales records from VGChartz, with 16,598 titles. The target is global sales in millions of units. Text features are game title and publisher name; structured features are platform (31 gaming systems), release year, and genre (12 categories).

Zomato Restaurants. Restaurant listings from the Zomato platform covering Bangalore, India, with 41,665 restaurants. The target is the aggregate user rating (ranging from 3.3 to 4.2). Text features include restaurant name, address, cuisine types, customer-highlighted dishes, raw user review text, and menu item lists; structured features include online ordering and table reservation availability, total votes, neighborhood location, restaurant type, and approximate cost for two.

E. Text-Tabular Curation

We evaluate existing text-tabular benchmarks by drawing candidates from 4 sources: the Multimodal AutoML Benchmark (Shi et al., 2021), Grinsztajn et al. (2023), CARTE (Kim et al., 2024), and TextTabBench (Mráz et al., 2025), yielding 56 unique candidates after deduplication and exclusion of datasets which were unavailable due to improper hosting. Each dataset is evaluated by 5 tabular learners over 5 folds under the 4 conditions defined in §2.2.

E.1. Existing Benchmarks

The 4 source benchmarks share a substantial number of datasets, either by directly using the exact same source or by using similar-enough datasets. We adopt the deduplication performed by Arazi et al. (2025), and extend it to include TextTabBench (Mráz et al., 2025), yielding a pool of 56 unique datasets. Table 3 shows datasets which are shared across more than one existing text-tabular benchmark.

Table 3. Duplicate datasets across benchmarks. ✓ indicates presence.

Dataset	AutoML Multimodal	Grinsztajn et al	CARTE	TextTabBench
Wine Reviews	✓	✓	✓	✓
Zomato Restaurants		✓	✓	
Vancouver Salaries		✓	✓	
Company Employees		✓	✓	
Montgomery Salaries		✓	✓	
Ramen Ratings		✓	✓	
Bike Bikewale		✓	✓	
Book Readability		✓	✓	
US Accidents		✓	✓	
Mercari Marketplace	✓			✓
California House Prices	✓			✓
Kickstarter Funding	✓			✓
Fake Job Posting	✓			✓
Spotify Genres		✓		✓
Beer Ratings			✓	✓

E.2. Empirical Results for Curation Conditions

Figure 5 shows normalized scores across all 4 conditions for the full pool and the MulTaBench subset. The *Structured* and *Unstructured* bars serve as unimodal baselines. The MulTaBench subset shows a consistent ordering across all 4 conditions, which is more pronounced than in the full corpus.

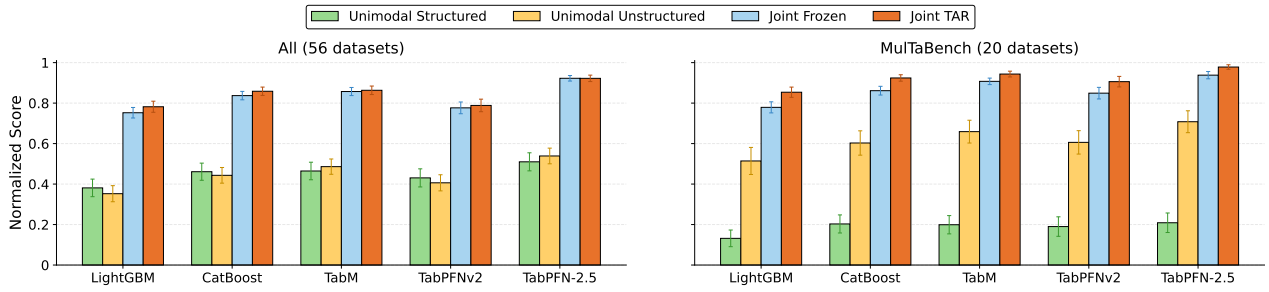

Figure 5. Curations Conditions for the Text-Tabular Pool. Normalized scores for *Structured*, *Unstructured*, *Joint Frozen*, and *Joint TAR* across all 56 candidates (left) and the MulTaBench subset (right).

E.3. Benchmark Acceptance Breakdown

Table 4 reports acceptance rates per source benchmark. Grinsztajn et al. and the AutoML Multimodal Benchmark yield the highest rates. CARTE has the lowest acceptance rate (33%), reflecting its focus on knowledge-graph-style short strings and high-cardinality categorical columns. Out of 56 candidates, 23 pass all criteria; we retain 20 for MulTaBench.

E.4. Per-Dataset Curation Results

Table 5 reports, for each of the 56 candidate datasets, whether each of the five curation models satisfies both criteria jointly. A checkmark indicates both hold simultaneously; × indicates failure on at least one; − indicates the model could not be evaluated, due to highly-multiclass problems where TabPFN's variants can't run on. Datasets are sorted approved-first, then by descending pass count.

Table 4. Text-tabular curation acceptance rates by source benchmark.

Benchmark	Candidates	Accepted	Rate
AutoML Multimodal	16	10	62%
Grinsztajn et al.	11	7	64%
CARTE	33	11	33%
TextTabBench	13	7	54%
Total (deduplicated)	**56**	**23**	**41%**

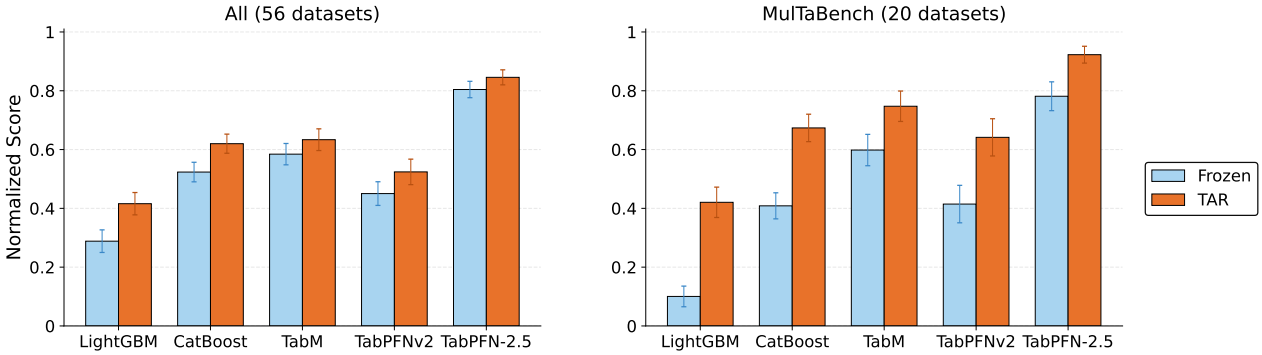

Figure 6. Target-Aware Representations Gains over Frozen. Normalized scores for *Joint TAR* and *Joint Frozen* across all text-tabular benchmark datasets (left), and its MulTaBench subset (right).

Table 5. Per-dataset curation grid. Models: LightGBM (LGBM), CatBoost (Cat), TabM, PFNv2 (TabPFNv2), PFN-2.5 (TabPFN-2.5). Each cell indicates whether the model satisfies criteria. *Pass?* column shows how many models pass. Options are pass (✓), fail (×), and N/A (−).

Dataset	LGBM	Cat	TabM	PFNv2	PFN-2.5	Pass?
Approved (23 datasets)						
Kickstarter Funding	✓	✓	✓	✓	✓	5
Jigsaw Toxicity	✓	✓	✓	✓	✓	5
Product Sentiment	✓	✓	✓	✓	✓	5
Women's Clothing	✓	✓	✓	✓	✓	5
Michelin Guide	✓	✓	✓	✓	✓	5
News Channel Category	✓	✓	✓	✓	✓	5
Baby Products	✓	✓	✓	✓	✓	5
Vancouver Salaries	✓	✓	✓	✓	✓	5
SciMagojr Impact	✓	✓	✓	✓	✓	5
Book Readability	✓	✓	✓	✓	✓	5
Video Games Sales	✓	✓	✓	✓	✓	5
Consumer Complaint	✓	✓	✓	✓	×	4
Hearthstone Cards	✓	✓	×	✓	✓	4
US Accidents	✓	✓	×	✓	✓	4
Book Price	✓	✓	✓	✓	×	4
Mercari Marketplace	✓	✓	✓	✓	×	4
Zomato Restaurants	✓	✓	✓	✓	×	4
Rotten Tomatoes	✓	✓	×	✓	✓	4
Fake Job Posting	✓	✓	×	✓	×	3
Wine Review	✓	✓	✓	−	−	3
Data Scientist Salary	×	✓	×	✓	✓	3
Spotify Genres	✓	✓	✓	−	−	3

(continued on next page)

(continued from previous page)

Dataset	LGBM	Cat	TabM	PFNv2	PFN-2.5	Pass?
Montgomery Salaries	×	×	✓	✓	✓	3
Rejected (33 datasets)						
OSHA Accident Injury	✓	×	✓	×	×	2
Google Q&A Type	×	×	×	✓	✓	2
American Eagle Prices	×	✓	✓	×	×	2
JC Penney Products	✓	×	×	✓	×	2
Wikiliq Alcohol	×	×	✓	✓	×	2
Chocolate Bar Ratings	×	✓	✓	×	×	2
Wine Vivino Spain	✓	×	✓	×	×	2
California House Prices	✓	✓	×	×	×	2
SF Permit Applications	×	×	✓	✓	×	2
FIFA22 Wages	✓	×	×	✓	×	2
IMDB Genre	×	×	×	✓	×	1
Melbourne Airbnb	✓	×	×	×	×	1
Bike Price Bikewale	×	×	×	✓	×	1
Car Price Cardekho	✓	×	×	×	×	1
Polish Wine Prices	×	✓	×	×	×	1
ML/DS Job Salaries	×	×	✓	×	×	1
Books Goodreads	×	✓	×	×	×	1
Korean Drama	✓	×	×	×	×	1
US Museum Revenues	✓	×	×	×	×	1
Used Cars Pakistan	✓	×	×	×	×	1
Used Cars Saudi Arabia	×	×	×	×	✓	1
Yelp Reviews	×	×	×	×	×	0
Laptop Indian Prices	×	×	×	×	×	0
Beer Ratings	×	×	×	×	×	0
Coffee Review	×	×	×	×	×	0
Ramen Ratings	×	×	×	×	×	0
Airbnb Seattle	×	×	×	×	×	0
Company Employee Size	×	×	×	×	×	0
Anime Planet Rating	×	×	×	×	×	0
FilmTV Movie Rating	×	×	×	×	×	0
Movies Dataset Revenue	×	×	×	×	×	0
NBA Draft VORP	×	×	×	×	×	0
Mercedes Italy Cars	×	×	×	×	×	0

F. Image-Tabular Curation

F.1. Existing Benchmarks

The image-tabular benchmarking landscape is substantially more limited than its text-tabular counterpart. MuG (Lu et al., 2023) reports 8 text-image-tabular datasets, but these correspond to only 4 underlying datasets, some of them using different target variables. Tang et al. (2024b) curate 22 datasets spanning varying modality combinations: 6 are text-tabular and overlap with text-tabular benchmarks; 5 are text-image datasets that lie outside the scope of this paper; and the remaining 11 qualify for our image-tabular definition (6 of them also have text). In addition, we include the datasets introduced by TIME (Luo et al., 2025b) and MultimodalTabPFN (Kim et al., 2025b), some of them overlapping with aforementioned benchmarks.

However, many of these datasets suffer from serious reproducibility problems. For example, the *Seattle* dataset contains links to images via external URLs that are no longer reachable; and the *KARD* dataset points to a Kaggle dataset that has since been deleted. The remaining candidates are partially recoverable, but their preprocessing logic is often undocumented

and difficult to replicate faithfully.

After deduplication and removal of unavailable datasets, we are able to evaluate 16 unique datasets, from which only 5 pass the curation filter. For the ones which did non pass, it was sometimes hard to assess whether we have curated them properly. Therefore, we do not report curation statistics at the same level of detail as for text-tabular, and focus the remaining of the section on elaborating on the curation process.

F.2. Curation Logic

The curation of datasets found in the wild involved several decisions with the goal of making the image feature important and interesting enough to qualify as a relevant true image-tabular task.

Images. Each dataset contains exactly one image column; datasets with multiple image fields per row (e.g., product galleries) were reduced to a single image for simplicity. Rows with absent or corrupt image files are dropped without imputation, as there is no sensible substitute for a missing image, and placeholder images would inject noise into the encoding step.

Feature and Target engineering. In several cases the raw target required transformation before satisfying the curation criteria, and we provide a non-exhaustive list of examples. *Log transformation*: Amazon Bestseller retail price is transformed as $\log(1 + \text{price})$ to stabilize the regression target across several orders of magnitude. *Quantile binning*: CS:GO Skin Price (10 equal-frequency bins), PetFinder listed age (8 bins), and HubMAP HPA donor age (10 bins) are discretized into multiclass targets. *Feature removal*: structured columns that directly encode the target or fully dominate the image signal are dropped. This is particularly evident in examples like Zooscan Plankton, where features were extracted directly from the image, and removing them increased the image importance.

Kaggle upload. To ensure reproducibility, all 20 image-tabular datasets are preprocessed and uploaded to Kaggle under the MulTaBench organization. Each upload contains a flat `images/` directory with one file per row named consistently, a `data.csv` with features and target. The image column stores relative paths into `images/` directory. A unified loading API handles download and ingestion, ensuring all datasets are accessed identically regardless of original source format.

G. Text-Image-Tabular Datasets

From the 20 image-tabular datasets in MulTaBench, 8 of them include one or more text columns alongside the image and structured features. To investigate whether this could be treated as true text-image-tabular datasets, we apply also the full text curation pipeline to each, by conducting the independent test elaborated on Appendix C.3 to both image and text. By applying the selection rule independently, we prove that the 3 modalities contribute to the prediction to fulfill the *Joint Signal* criterion. In addition, for *Task-awareness*, we require that TAR on the image and on the text would improve on the respective frozen conditions. Finally, we also explicitly demand that performing TAR over both modalities (i.e., finetuning both the image and text encoder, separately) would improve on finetuning only one of them.

Of the 8 candidates, we find that only two satisfy all criteria for at least 3 learners: *PetFinder* and *Amazon Packages*. The remaining 6 fail primarily because text TAR does not improve over the frozen joint baseline, which might be a relative strict requirement. For future text-image-tabular efforts, one could consider relaxing this last condition, by only demanding that at least one of the modalities gains from representation tuning.

Table 6. The PetFinder Analysis. S=Structured, I=Image, T=Text. For all models, performing Joint Modeling and Target-Aware Representations for both modalities maximizes AUC (shown in %).

	Single modality		*Frozen combinations*			*Target-Aware Representations (TAR)*		
Model	I	T	S+I	S+T	S+I+T	S+I_{TAR}+T	S+I+T_{TAR}	**S+I_{TAR}+T_{TAR}**
LightGBM	77.2	72.1	79.9	77.7	81.1	82.8	84.2	**85.7**
CatBoost	78.9	73.5	81.7	79.3	83.2	83.9	85.2	**86.4**
TabM	80.2	74.9	83.0	80.7	84.2	84.8	86.3	**87.0**
TabPFNv2	80.7	73.5	83.2	79.3	83.9	84.5	86.3	**87.1**
TabPFN-2.5	81.1	76.0	83.7	81.0	84.9	85.3	87.3	**88.0**

Table 6 presents a case study of this analysis for the *PetFinder* dataset., and Table 7. does the same for the *Amazon Packages* dataset.

Table 7. Amazon Packages Analysis. S=Structured, I=Image, T=Text. Mean R^2 (%) per model and condition. For all models, TAR over both modalities dominates.

	Single modality		*Frozen combinations*			*Target-Aware Representations (TAR)*		
Model	I	T	S+I	S+T	S+I+T	S+I_{TAR}+T	S+I+T_{TAR}	**S+I_{TAR}+T_{TAR}**
LightGBM	43.2	17.5	45.4	20.8	49.9	56.9	52.2	**59.4**
CatBoost	44.2	19.0	46.7	22.5	52.5	58.1	53.8	**59.8**
TabM	46.6	20.5	48.6	24.2	55.7	60.3	56.1	**61.1**
TabPFNv2	46.6	20.2	49.3	23.9	54.9	59.8	56.2	**61.3**
TabPFN-2.5	47.5	21.1	50.2	24.9	56.2	60.7	57.2	**61.5**

H. Extended Results

H.1. Main Results Breakdown

New Models. We extend our models suite by adding new models.

- For XGBoost, we follow previous work (Gorishniy et al., 2021; Arazi et al., 2025) and use the default implementation from the *xgboost* package,[7] with $booster = gbtree, early\_stopping\_rounds = 50, n\_estimators = 2000$.
- For RandomForest, we use the default scikit-learn implementation with default configuration with $n\_estimators = 100$.
- For *RealMLP* we use its official implementation in the *pytabkit* package, disable label smoothing and optimize for cross_entropy for binary classification and $1 - auc\_ovr$ for multiclass classification, keeping the other default hyperparameters.
- For *TabICLv2,*[8] *TabSTAR,*[9] *ConTextTab,*[10] and *TabDPT,*[11] we use their default implementations.
- For *AutoGluon-Multimodal* we use *MultiModalPredictor*[12] with $pretrained = True$ and optimizing for roc_auc (binary classification), roc_auc_ovr (multiclass classification), and r^2 (regression).

Task Type Breakdown Figures 7 and 8 replicate Figure 2, but breaking down to classification and regression datasets respectively. TAR consistently outperforms Frozen in both task types and both modalities, indicating that the benefit of target-aware representations is not specific to any of them.

Win Rate by Model Table 8 reports the fraction of (dataset, fold) pairs where TAR outperforms Frozen for each model, with 95% CIs. End-to-end systems that do not expose a separate TAR condition for a given modality (TabSTAR, ConTextTab) are excluded from the corresponding column. TAR beats Frozen in the large majority of runs across all models and both modalities.

Per-dataset Results. Tables 9 and 10 report per-dataset results for all 20 image-tabular and 20 text-tabular datasets, averaged over all learners that have both Frozen and TAR conditions and 5 random seeds, sorted by TAR gain. Negative R^2 scores are clipped before averaging.

[7] https://pypi.org/project/xgboost/
[8] https://pypi.org/project/tabicl/
[9] https://pypi.org/project/tabstar/
[10] https://github.com/SAP-samples/sap-rpt-1-oss
[11] https://pypi.org/project/tabdpt/
[12] https://auto.gluon.ai/stable/api/autogluon.multimodal.MultiModalPredictor.html

Table 8. Per-model TAR win rate on MulTaBench. End-to-end models excluded from columns where they lack a separate TAR condition.

Model	Image (%)	Text (%)	All (%)
CatBoost	90.0 ± 5.9	93.0 ± 5.0	91.5 ± 5.5
LightGBM	84.0 ± 7.2	93.0 ± 5.0	88.5 ± 6.2
RF	85.0 ± 7.0	90.0 ± 5.9	87.5 ± 6.5
XGBoost	80.0 ± 7.8	90.0 ± 5.9	85.0 ± 6.9
TabPFNv2	77.0 ± 8.2	84.4 ± 7.5	80.7 ± 7.9
TabM	82.0 ± 7.5	77.0 ± 8.2	79.5 ± 7.9
TabDPT	70.0 ± 9.0	87.0 ± 6.6	78.5 ± 7.9
RealMLP	78.0 ± 8.1	78.0 ± 8.1	78.0 ± 8.1
TabSTAR	76.0 ± 8.4	—	76.0 ± 8.4
TabPFN-2.5	77.0 ± 8.2	73.3 ± 9.1	75.2 ± 8.7
ConTextTab	73.0 ± 8.7	—	73.0 ± 8.7
TabICLv2	55.0 ± 9.8	75.0 ± 8.5	65.0 ± 9.2

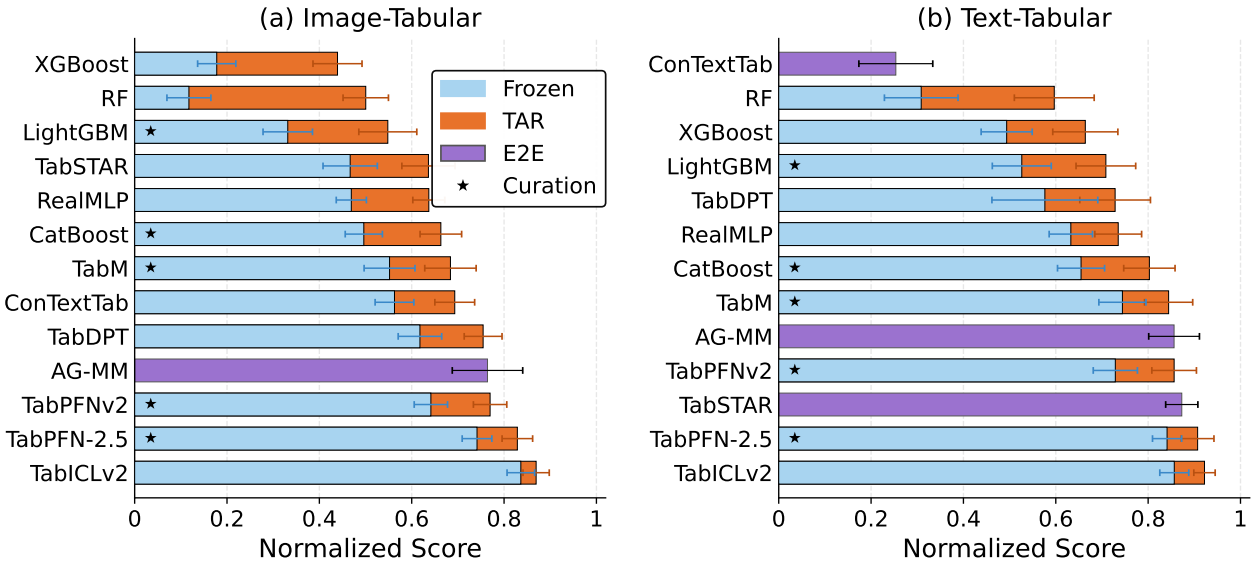

Figure 7. Tabular Learners Performances Analysis for Classification Tasks. Normalized scores over MulTaBench, with ± 95% CI.

H.2. Missing Baselines

We deliberately exclude autoregressive generative models (LLMs and VLMs) from the benchmark evaluation, due to prohibitive inference costs and a memorization risk. The research on benchmarking LLMs and VLMs for MMTL task still needs to be explored. Although TIME (Luo et al., 2025b) and MultimodalTabPFN (Kim et al., 2025b) are relevant baselines, TIME has not released the code at the time of our submission. MultimodalTabPFN, on the contrary, has a working codebase, but it is highly not flexible to serve the model using the popular *scikit-learn* (Pedregosa et al., 2011) wrapper, making it hard to evaluate.

H.3. Computation Costs

Table 11 and Figure 9 reports median wall-clock runtimes and peak GPU memory per (dataset, fold) run on a single *NVIDIA A100-SXM4* GPU with 40GB memory, and 8 CPU cores of type *AMD EPYC 7742 64-Core Processor*. We report results for each of the 5 core learners across frozen and TAR conditions and both encoder sizes. From the table, it is evident that the embeddings dominate all metrics. TAR adds a substantial overhead relative to frozen embeddings, dominated by the encoder fine-tuning step. For image datasets with the small DINO encoder, TAR roughly doubles runtime; the large encoder raises costs further.

Text TAR is significantly more expensive: *e5-small* TAR takes roughly ten times longer than frozen, and *e5-large* TAR approaches three hours per run. The gap arises partly as text-tabular datasets often contain more than a single text column,

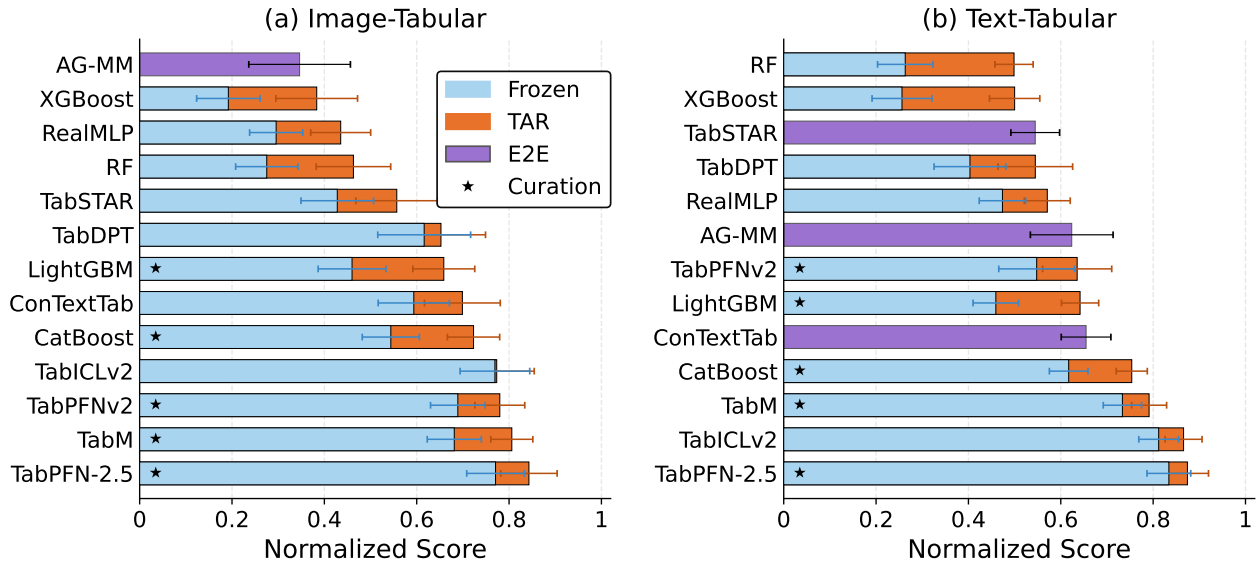

Figure 8. Tabular Learners Performances Analysis for Regression Tasks. Normalized scores over MulTaBench, with ± 95% CI.

making their effective dataset size much bigger.

The costs above are measured without any hyperparameter optimization (HPO). Standardizing HPO across 40 datasets is computationally prohibitive under the TAR paradigm: the encoder must be fine-tuned separately for each cross-validation fold to prevent data leakage, so a standard HPO sweep would require repeating encoder fine-tuning for every hyperparameter trial, multiplying an already expensive operation by the number of trials. Consequently, all experiments use a single fixed LoRA configuration across all datasets, with no per-dataset tuning of the encoder or the learner. All reported gains should therefore be interpreted as conservative lower bounds on what a fully tuned system could achieve.

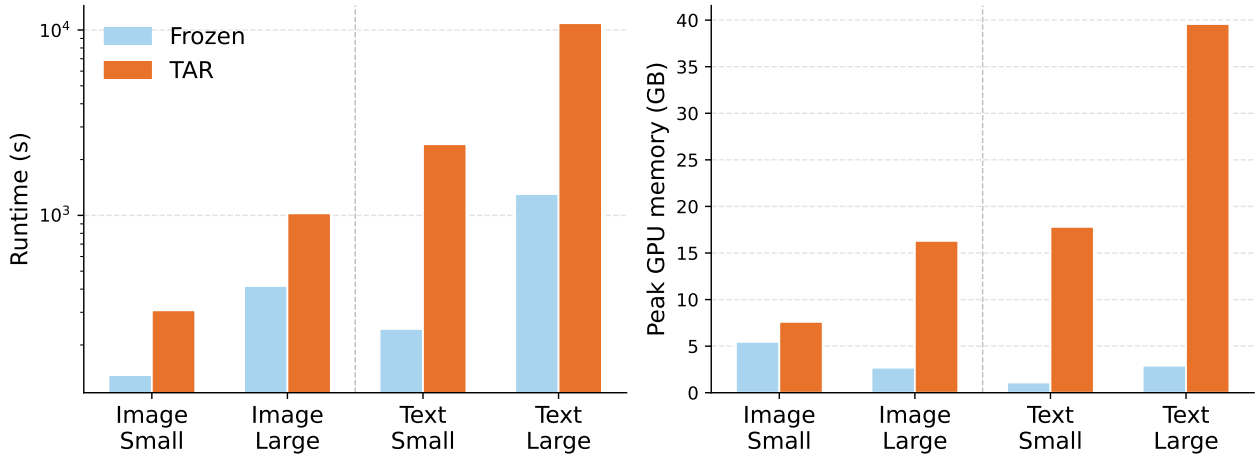

Figure 9. Computation costs per run. Left: median runtime in seconds (log scale). Right: median peak GPU memory. The dashed vertical line separates image (left) and text (right) conditions.

H.4. Encoder Scale by Task Type

We replace our default encoders, DINO-small and E5-small, with DINO-large[13] and e5-large. Roughly speaking, this moves from models of 30M paremeters to models of 300M parameters. We then re-evaluate all 20 image and 20 text datasets.

[13]The official names are *dinov3-vits16-pretrain-lvd1689m* for small, and *dinov3-vitl16-pretrain-lvd1689m* for large.

Table 9. MulTaBench Image-Tabular Per-dataset Results. Averaged over 12 learners and 5 seeds, with both Frozen and TAR conditions, sorted by Gain. AUROC for classification, R^2 for regression.

Dataset	Frozen	TAR	Gain
Mango Mass	0.533	0.653	+0.120
Khaadi Clothes	0.565	0.683	+0.118
Amazon Packages	0.523	0.579	+0.056
CheXpert	0.762	0.803	+0.041
CBIS-DDSM	0.858	0.893	+0.035
Amazon Bestseller	0.543	0.566	+0.024
MkPhoto Bots	0.341	0.361	+0.020
Justin Instagram	0.956	0.974	+0.017
Hateful Meme	0.745	0.760	+0.015
H&M Fashion	0.389	0.404	+0.015
Glaucoma SMDG	0.921	0.934	+0.012
Mammography CMMD	0.770	0.780	+0.010
CelebA Attractiveness	0.903	0.913	+0.009
PetFinder	0.832	0.841	+0.009
HubMAP HPA	0.959	0.968	+0.009
Flower Bouquets	0.627	0.636	+0.009
Painting Price	0.270	0.275	+0.005
Letterboxd Movies	0.439	0.443	+0.004
Zooscan Plankton	0.983	0.985	+0.003
CS:GO Skins	0.871	0.870	−0.001
Mean	0.682	0.703	+0.022

Figures 11 and 12 replicate Figure 10 restricted to classification and regression datasets respectively. TAR consistently outperforms Frozen across encoder sizes and task types, confirming that the benefit of TAR generalizes also to larger encoders.

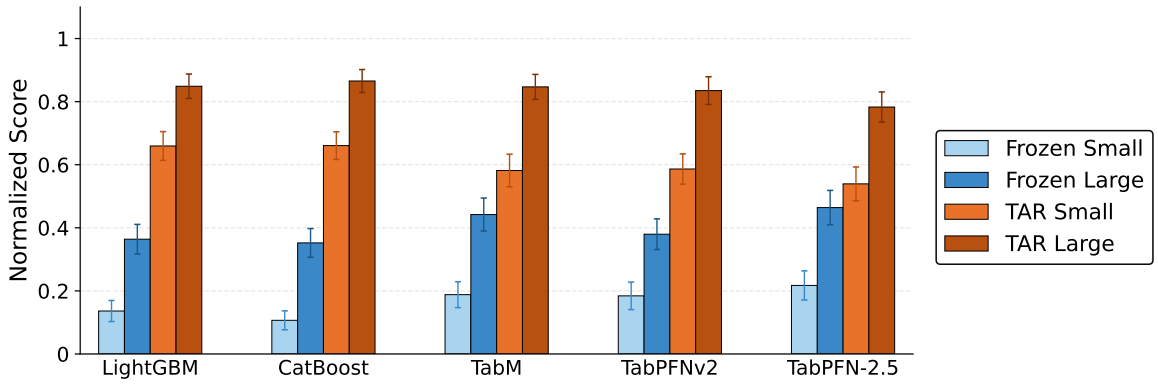

Figure 10. Embedding Model Size Analysis. Normalized scores are computed with min-max scaling at the learner level. TAR variants outperform the frozen ones for both model sizes.

H.5. PCA Dimensions Analysis

To verify that the gains from target-aware adaptation do not depend on the PCA compression step, we repeat the core Frozen vs. TAR comparison using raw 384-dimensional embeddings, omitting the projection entirely. Since this largely increases the number of features for the downstream task, we limit the analysis to CatBoost and LightGBM, and exclude datasets with more than 5 text features, resulting in 33 datasets. Figure 14 shows 4 conditions side by side, varying between N=30 and No-PCA, and Frozen vs TAR. We observe that TAR outperforms Frozen in both settings, for both learners, confirming that the advantage is not an artifact of dimensionality reduction. The signal surfaced by fine-tuning is present in the raw 384-dimensional space and persists regardless of whether embeddings are subsequently compressed.

Table 10. MulTaBench Text-Tabular Per-dataset Results. Averaged over 10 learners and 5 seeds, with both Frozen and TAR conditions, sorted by Gain. AUROC for classification, R^2 for regression.

Dataset	Frozen	Contextualized	Gain
Jigsaw Toxicity	0.806	0.926	+0.119
Video Games Sales	0.348	0.385	+0.036
Mercari	0.412	0.436	+0.024
Baby Products	0.873	0.895	+0.022
Vancouver Salaries	0.753	0.775	+0.022
Rotten Tomatoes	0.490	0.508	+0.018
Kickstarter	0.720	0.737	+0.017
Book Price	0.529	0.543	+0.013
Zomato Restaurants	0.818	0.832	+0.013
Michelin Guide	0.896	0.909	+0.013
Book Readability	0.809	0.821	+0.012
Wine Review	0.968	0.976	+0.009
Product Sentiment	0.901	0.909	+0.009
SciMagojr Impact	0.852	0.860	+0.009
US Accidents	0.965	0.974	+0.008
Women's Clothing	0.903	0.909	+0.006
Spotify Genres	0.935	0.940	+0.005
Data Scientist Salary	0.823	0.828	+0.005
Montgomery Salaries	0.968	0.972	+0.004
Fake Job Postings	0.916	0.918	+0.002
Mean	0.774	0.792	+0.018

I. Additional Attention Maps

Each of the 4 datasets in Figure **??** is accompanied by 3 additional test-set examples below. In every case, Frozen attention remains scattered across task-irrelevant regions, while Target-Aware attention converges on semantically meaningful area identified in the main figure, relevant to the prediction.

J. Discussion and Conclusion

In this work, we introduce MulTaBench, a benchmark of 40 image-tabular and text-tabular datasets designed to explore challenging Multimodal Tabular Learning tasks. We contribute the largest image-tabular benchmark to date, while focusing on tasks that benefit from Joint Modeling and TAR, differing ourselves from existing MMTL benchmarks. Our findings show that existing models rely on representations that are often insufficient for the task at hand, making MulTaBench a necessary tool for evaluating the next generation of Multimodal Tabular Foundation Models.

MulTaBench suffers from an important limitation: our curation pipeline entangles the computational problem with the algorithmic solution. As such, it is hard to predict in advance whether a new dataset meets our criteria, and the models used for the curation cannot be fairly evaluated due to selection bias. While we believe future research should aim to address these limitations, our work is a strong step to tackle a problem which was overlooked so far, yielding findings that generalize well to new models. Importantly, the automated nature of our pipeline facilitates the continuous expansion of MulTaBench to include new dataset candidates, the latest tabular learners, or refined selection logic as the field matures. As such, our curation pipeline is a contribution of its own, providing a mechanism to refresh the benchmark with harder candidates as current tasks become saturated by future models.

Our research paves the way to many exciting future directions, such as expanding to a dedicated text-image-tabular benchmark, exploring other modalities such as audio and videos, or analyzing different prompting strategies to steer embeddings towards the target. Mainly, MulTaBench supports the development of Multimodal TFMs. In our opinion, there are two big challenges to solve: architecture and training data. For architectures, in §5, we suggest that future models should ideally take the best out of ICL and finetuning; for instance, coupling TFMs with LLMs and VLMs is a compelling path. For training data, since real data corpora for MMTL are rare, (Eggert et al., 2023), expanding the syntethic numerical priors used for training TFMs (Hollmann et al., 2025; Zhang et al., 2025a; Qu et al., 2026) to include text and image features is an exciting direction (Luo et al., 2025a; Brahmavar et al., 2026). We hope that our work will contribute to the research of Multimodal Tabular Learning, and we are excited towards a future where this crucial problem sees the progress it deserves.

Table 11. Computation costs runs. Median Runtime in seconds and Median Peak GPU memory in GB. Partition by tabular learners, modality and encoder size.

	Small Encoder				*Large Encoder*			
	Runtime (s)		Peak GPU (GB)		Runtime (s)		Peak GPU (GB)	
Model	Frozen	TAR	Frozen	TAR	Frozen	TAR	Frozen	TAR
Image-Tabular (DINO encoder)								
LightGBM	141	287	5.5	7.6	411	1,003	2.7	16.7
CatBoost	128	307	5.4	7.6	432	1,002	2.7	16.7
TabM	129	282	5.4	7.6	375	993	2.7	16.8
TabPFNv2	173	318	5.3	7.6	432	1,055	2.7	14.8
TabPFN-2.5	169	316	5.5	7.6	405	1,048	2.7	14.8
Text-Tabular (E5 encoder)								
LightGBM	223	2,417	1.1	17.3	1,284	10,867	2.9	39.6
CatBoost	323	2,471	1.1	17.3	1,377	10,860	2.9	39.6
TabM	225	2,408	1.2	17.3	1,256	10,833	2.9	39.6
TabPFNv2	253	2,430	1.1	18.6	1,368	11,569	2.9	39.6
TabPFN-2.5	242	2,396	1.1	18.6	1,347	11,556	2.9	39.6

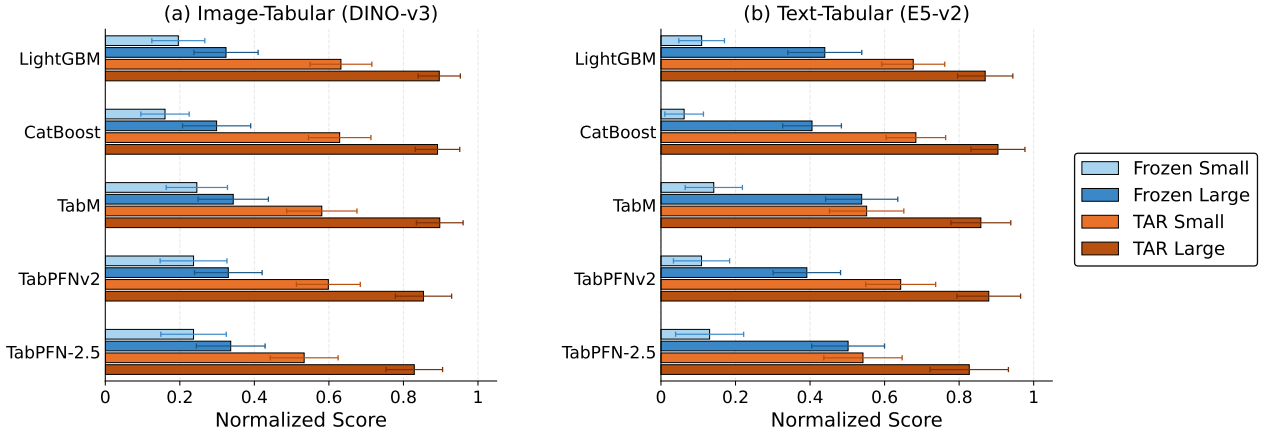

Figure 11. Encoder Scale Analysis for Classification. Small and large encoder variants, frozen and TAR, normalized within each model.

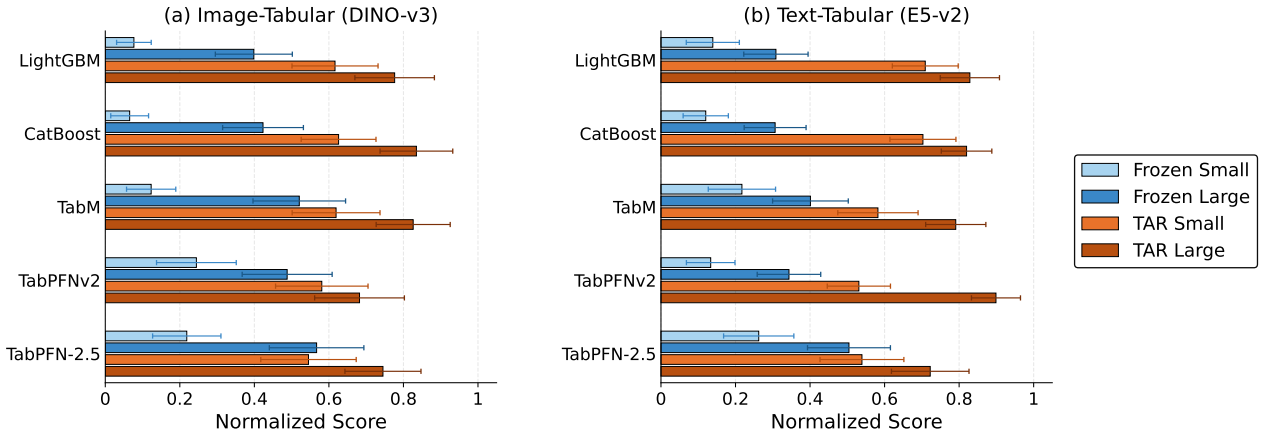

Figure 12. Encoder Scale Analysis for Regression. Small and large encoder variants, frozen and TAR, normalized within each model.

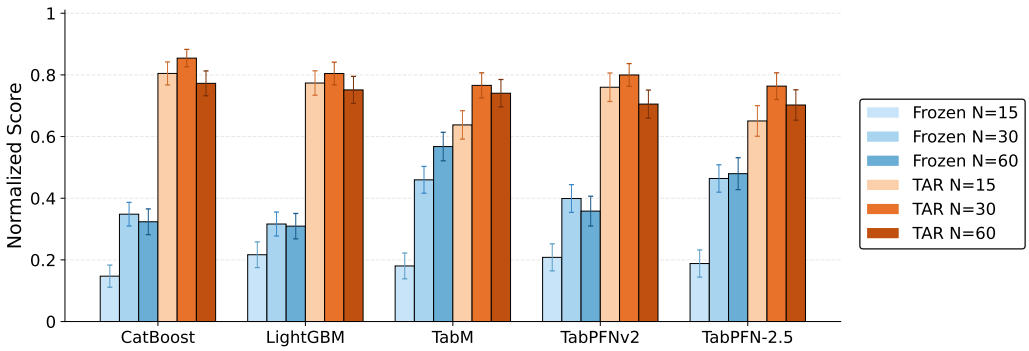

Figure 13. Embedding Dimension Analysis. Normalized scores are computed with min-max scaling at the learner level. TAR variants are stronger than Frozen ones for 15, 30, and 60 PCA components.

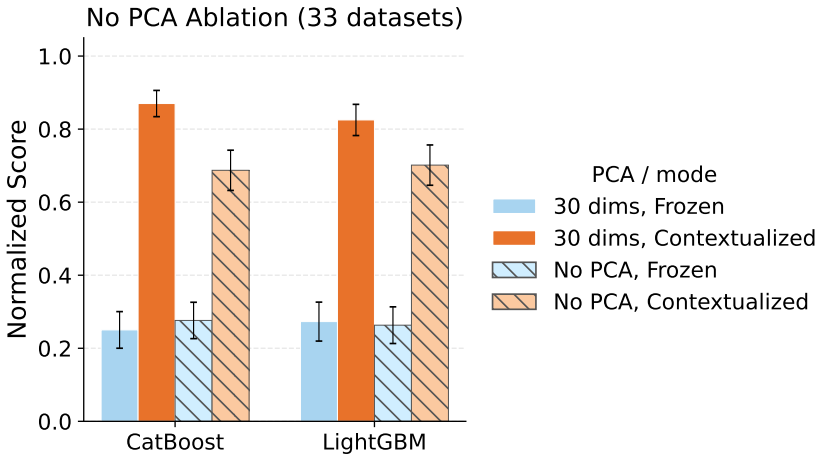

Figure 14. No-PCA ablation on 33 datasets for CatBoost and LightGBM. Normalized scores are on the model level.

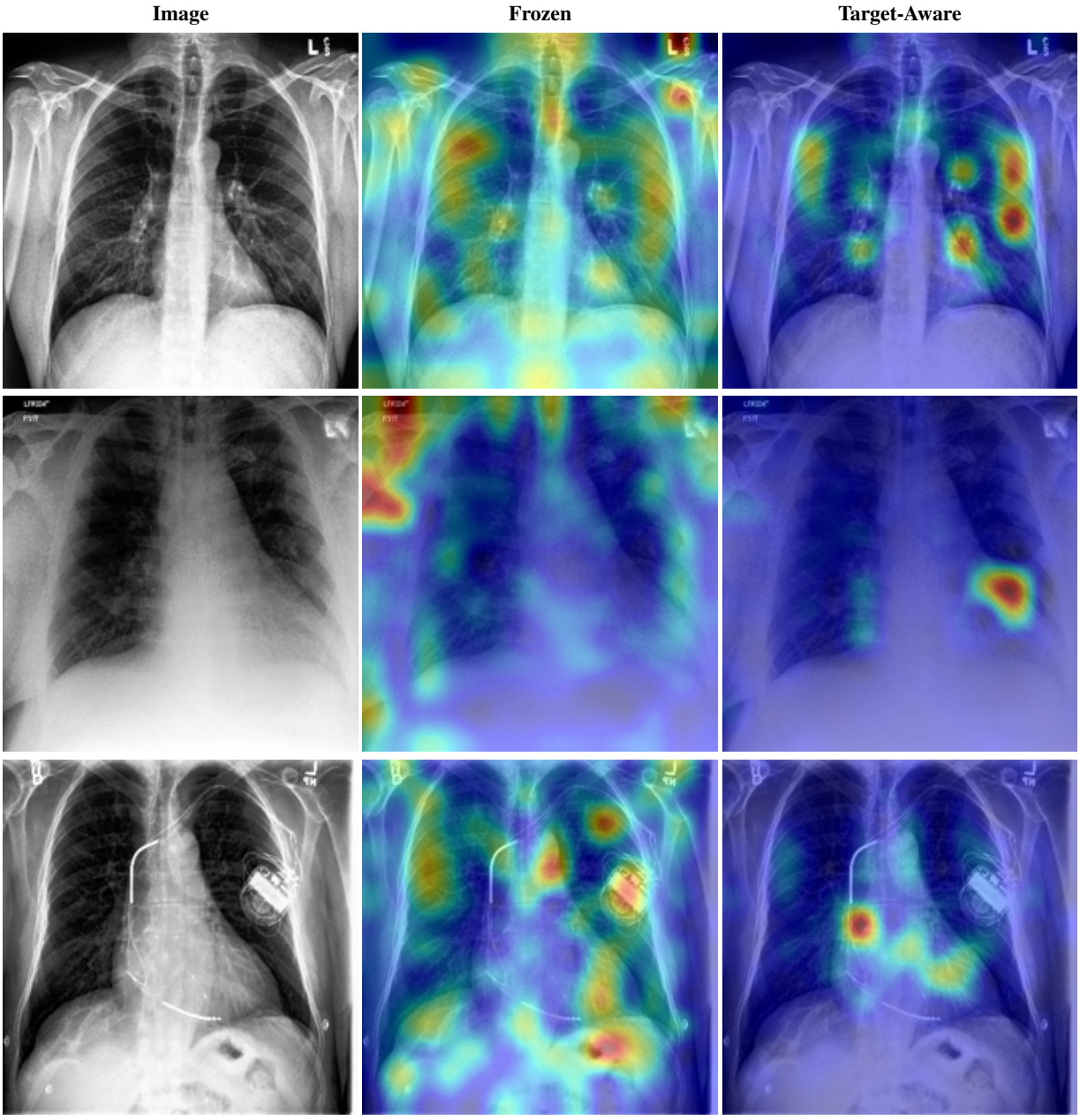

Figure 15. CheXpert Attention Maps. The attention shifts from diffused edges to the lung.

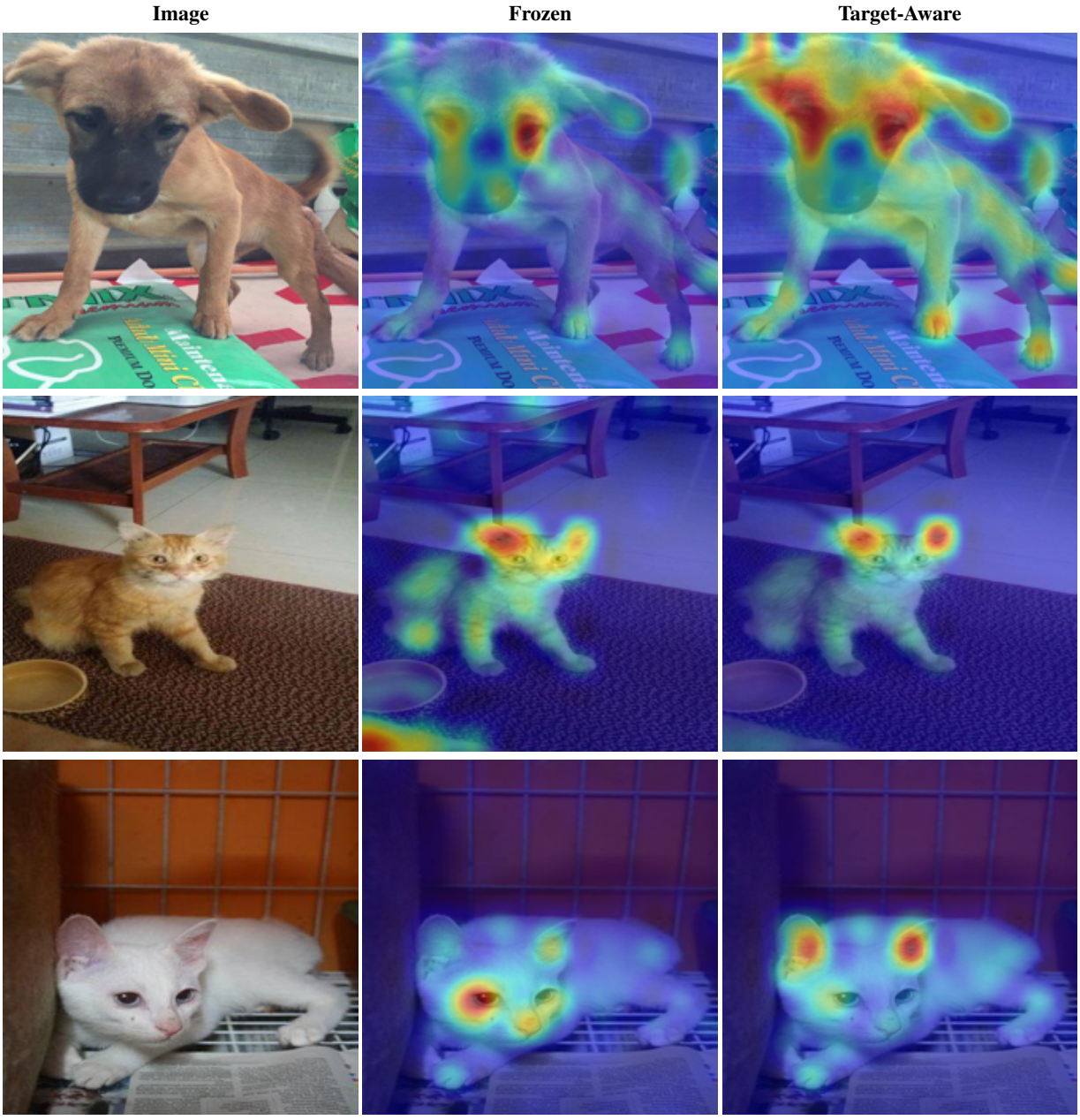

Figure 16. PetFinder Attention Maps. Attention isolates the cat ears and the dog's eyes.

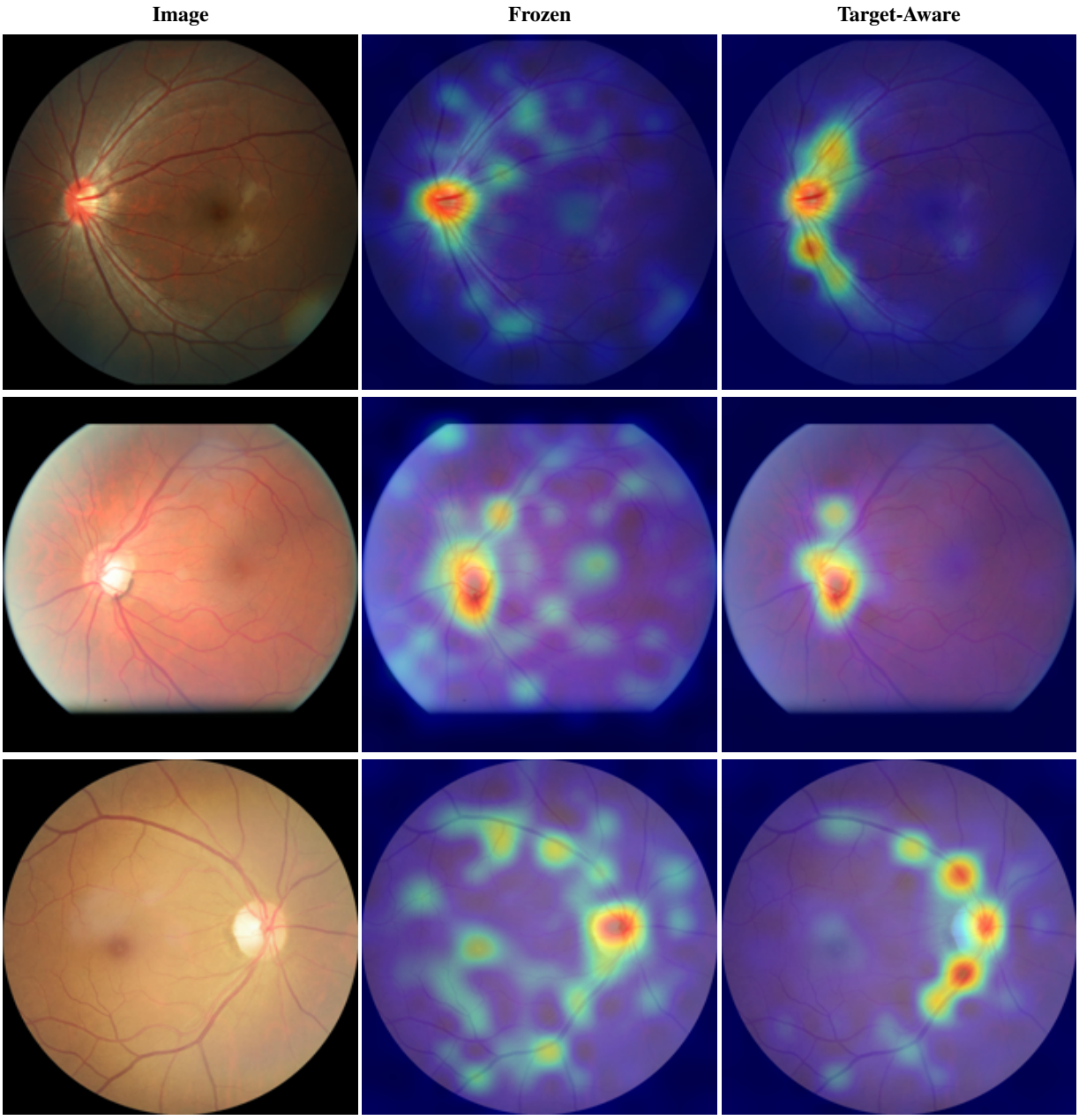

Figure 17. Glaucoma Attention Maps. Frozen attention scatters randomly across the retina; TAR converges on the optic disc and nerve fiber region, the clinically relevant area for glaucoma diagnosis.

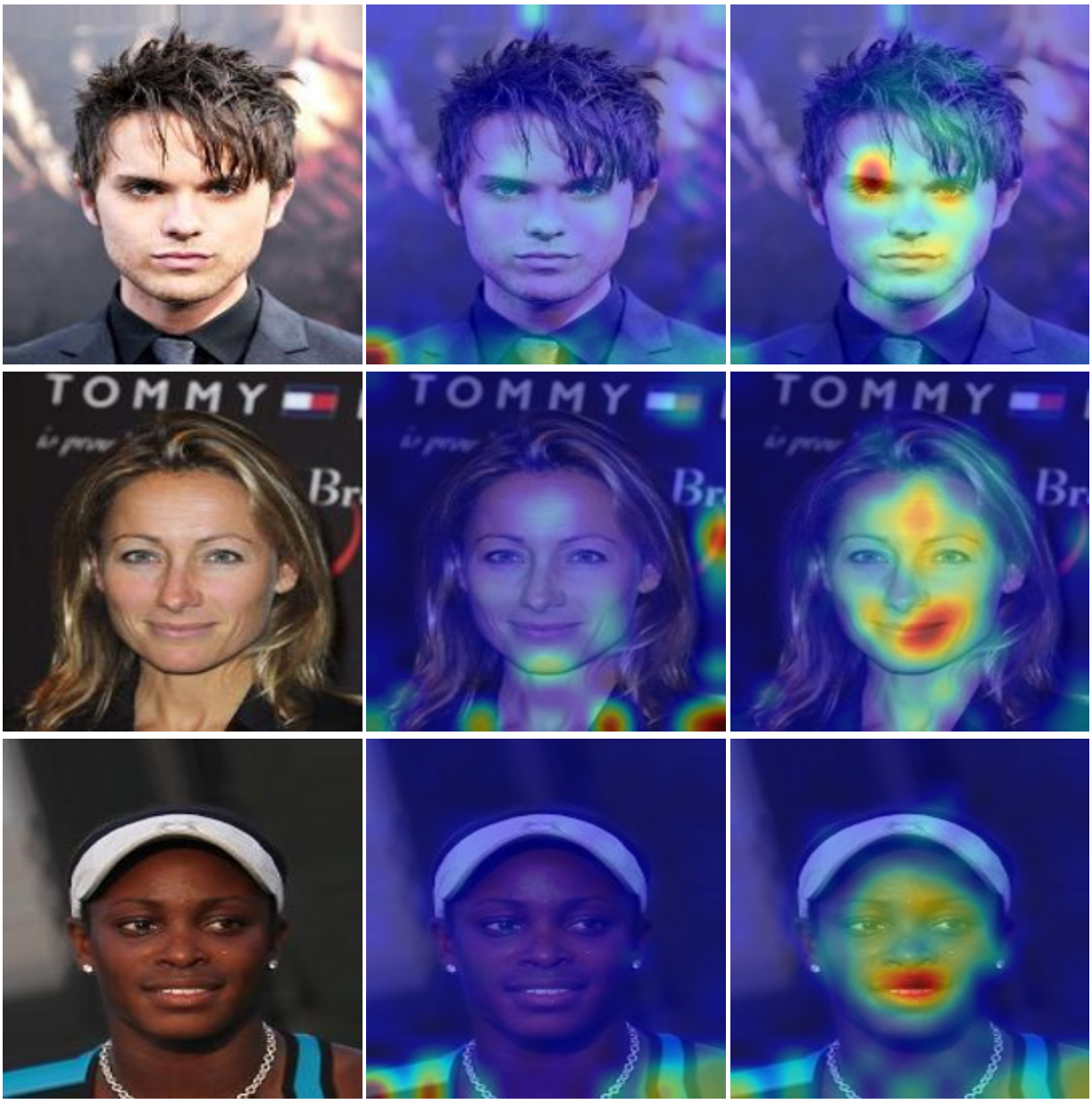

Figure 18. Celeb Attractiveness Attention Maps. Frozen attention disperses across accessories, clothing, and background; TAR consistently focuses on facial features.