

Harmonized Tabular-Image Fusion via Gradient-Aligned Alternating Learning

Longfei Huang¹, Yang Yang^{1*}

¹Nanjing University of Science and Technology

{hlf, yyang}@njust.edu.cn

Abstract—Multimodal tabular-image fusion is an emerging task that has received increasing attention in various domains. However, existing methods may be hindered by gradient conflicts between modalities, misleading the optimization of the unimodal learner. In this paper, we propose a novel Gradient-Aligned Alternating Learning (GAAL) paradigm to address this issue by aligning modality gradients. Specifically, GAAL adopts an alternating unimodal learning and shared classifier to decouple the multimodal gradient and facilitate interaction. Furthermore, we design uncertainty-based cross-modal gradient surgery to selectively align cross-modal gradients, thereby steering the shared parameters to benefit all modalities. As a result, GAAL can provide effective unimodal assistance and help boost the overall fusion performance. Empirical experiments on widely used datasets reveal the superiority of our method through comparison with various state-of-the-art (SoTA) tabular-image fusion baselines and test-time tabular missing baselines. The source code is available at <https://github.com/njustkmg/ICME26-GAAL>.

Index Terms—Tabular-image Fusion, Multimodal Learning, Cross-modal Interaction, Gradient Conflict

I. INTRODUCTION

In recent years, tabular data is increasingly accessible in multimodal datasets, and its integration is crucial in various applications [1]–[3]. An emerging example is tabular-image fusion that involves integrating structured tables and images to provide a holistic understanding of subjects. Despite numerous achievements, existing tabular-image fusion approaches [4], [5] often follow multimodal joint learning, where gradient conflicts frequently arise. These modality gradient conflicts are primarily caused by the unified optimization objective in joint learning [6], [7]. This potentially misleads unimodal learning and results in suboptimal final performance. As shown in Figure 1(a), we visualize the cosine similarity between image gradients and multimodal gradients on DVM [8] dataset. Negative cosine similarity indicates the presence of conflicts.

Fortunately, recent multimodal studies [9]–[11] have recognized this issue and attempted to provide solutions. An early representative work, OGM [12], balances strong and weak modalities by adjusting gradient magnitudes. Subsequent work, MMPareto [11], introduces Pareto methods from multi-task learning to regulate gradient directions. Beyond the above joint learning paradigm, other works [13]–[15] adopt an alternating learning paradigm to decouple combined gradient into unimodal gradients, achieving higher performance. To facilitate interaction, they adopt a shared-head strategy.

*Corresponding author.

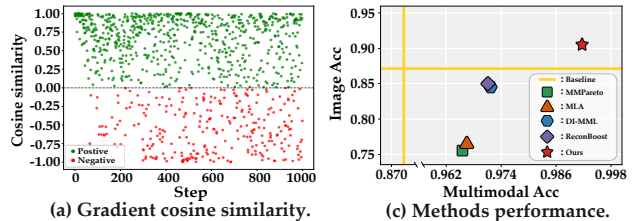


Fig. 1. We visualize the gradient conflicts and evaluate the performance of existing multimodal solutions on DVM dataset. (a) Multimodal and image gradients often show negative cosine similarity in naive joint learning, indicating severe gradient conflicts. (b) Existing multimodal solutions underperform in tabular-image fusion, failing to fully exploit unimodal image potential.

However, the aforementioned methods mainly focus on alleviating gradient conflicts in unimodal encoders, at the expense of the facilitative role of gradient signals in cross-modal interaction layer. This issue becomes pronounced when directly transferred to more challenging tabular-image tasks. As shown in Figure 1(b), these methods suffer from significant performance degradation in the tabular-image fusion, with image performance even falling below that of unimodal learning. This can be attributed to two main issues. First, gradient modulation is inherently more challenging in joint learning, making it difficult to unleash unimodal potential effectively. Second, while alternating learning can liberate unimodal encoders, shared head suffers from gradient conflicts. MLA [13] uses gradient orthogonalization to encourage modal independence, neglecting overlaps and synergy among modal objectives.

In this work, to overcome the above challenges, we propose Gradient-Aligned Alternating Learning (GAAL), a novel tabular-image learning paradigm that effectively addresses gradient conflict and strengthens the tabular-image fusion performance. Specifically, GAAL alternates modality-specific optimization to decouple multimodal gradients while capturing cross-modal interaction via a shared classifier, thereby performing relatively independent unimodal training. Then, we design uncertainty-based cross-modal gradient surgery that modulates modal gradients via quadratic programming (QP) [16], [17] to address gradient conflicts and synergy at shared classifier. This surgery computes gradients from high-entropy samples of the previous modality and projects the current modality gradient onto the previous gradient direction while minimizing their Euclidean distance. To this end, GAAL addresses gradient conflicts and facilitates interactions in tabular-

image fusion. As shown in Figure 1(b), our proposed method GAAL improves both multimodal and unimodal performance. Our main contributions are outlined as follows:

- We propose GAAL, a novel alternating learning algorithm that allows the model to explore unimodal information effectively, thereby improving tabular-image fusion.
- We design an uncertainty-based cross-modal gradient surgery that guides modal optimization directions to address gradient conflicts and facilitate synergy at the interaction layer.
- Experimental results demonstrate that GAAL effectively addresses gradient conflict to improve tabular-image fusion performance and achieves SoTA performance.

II. RELATED WORK

A. Tabular-image Fusion

Tabular data, known for its structured and interpretable nature, is the focus of traditional machine learning and deep learning [18], [19]. Unlike free-form audio or textual data, structured tabular data mixes dense numerical and sparse categorical features with differing value ranges and semantics, and lacks clearly defined interrelationships [20]. Existing tabular-image fusion approaches focus on integrating cross-modal information by multimodal learning strategies. TIP [5] proposes a transformer-based table encoder and enhances the robustness of tabular-image models through contrastive learning and cross-attention fusion. Moreover, since obtaining real-world tabular data is often costly and challenging, other studies have explored settings where tabular data are unavailable at inference time. These methods transfer expensive tabular expertise to images during training to enhance the performance of image models at inference. Since our method can facilitate cross-modal interaction, it also maintains good performance even when test-time tabular data are unavailable.

B. Multimodal Gradient Conflict

Gradient conflict, defined as negative cosine similarity between gradients, has been extensively studied in multiple domains [21], [22]. Recent studies [9], [11] have found that gradient conflicts also exist in multimodal scenarios. Such multimodal gradient conflicts can mislead unimodal optimization, ultimately resulting in suboptimal multimodal performance. To address this issue, some early works [9], [12] adjust the magnitude of modality-specific gradients based on unimodal performance to compensate for weak modality. Subsequent work MMPareto [11] introduces Pareto approach in multi-task learning to identify integrated gradient directions beneficial for all modalities. Moreover, since multimodal gradient conflicts mainly arise from the unified optimization objective in joint learning, recent methods adopt an alternating learning paradigm to decouple optimization objective. Although this framework can directly address gradient conflicts in the encoder, shared interaction layers across modalities still suffer from gradient conflicts. Notably, MLA also considers interaction layer, but it isolates modal gradients through orthogonalization, overlooking potential overlap and synergy among

modal optimization objectives. Therefore, our method aims to address both gradient conflicts and cross-modal interaction.

III. METHODOLOGY

In this section, we introduce the proposed GAAL, a tabular-image alternating learning framework that addresses gradient conflict to improve classification. To facilitate tabular-image interaction, we design an uncertainty-guided cross-modal gradient surgery, which utilizes uncertainty-based cross-modal gradient guidance to assist in optimizing unimodal samples. The architecture of GAAL is shown in Figure 2.

A. Preliminary

Let $\mathcal{X} = \{x_i^I, x_i^T, y_i\}_{i=1}^N$ be a training set, where y_i is the label for the i -th instance and N is the number of training data. $x_i^I \in \mathbb{R}^{H \times W \times 3}$ represents the image, and $x_i^T \in \mathbb{R}^D$ represents the tabular description, where D is the number of tabular features. Each tabular input contains two kinds of attributes, i.e., the categorical features (such as the ‘‘Gender’’) and the continuous features (such as the ‘‘Age’’). Assume there are Y classes in total, and $y_i \in [Y] = \{1, \dots, Y\}$. For the sake of simplicity, we use superscript m to indicate the module corresponding to a specific modality in this section, where $m \in \{I, T\}$. Notation is summarized in supplemental material.

B. Gradient-Aligned Alternating Learning

With the rapid growth of deep learning, representative tabular-image fusion approaches [5], [23], [24] have adopted deep neural network (DNN) for multimodal learning. Following these methods, we utilize DNN to construct our tabular and image models. Specifically, We use ϕ^m as encoders to extract features $u^m = \phi^m(\theta^m, x^m)$, where θ the encoder parameters. We define the predictor ψ as a mapping from the latent feature space to the label space. For given tabular-image pairs $x = [x^I, x^T]$, the tabular-image model can be written as $f(x) = \psi([u^I : u^T])$. Therefore, the objective function can be written as:

$$\mathcal{L}(x, y) = -\frac{1}{N} \sum_{i=1}^N y_i^\top \log(f(x)_i). \quad (1)$$

For a given input data instance, the alternating paradigm selects and updates a specific modality learner $f^m(x^m) = \psi(\Theta^m, \phi^m(x^m))$ at each time, where ψ is a shared classifier and Θ denotes the parameters of the classifier. The objective function $\mathcal{L}^m(x^m, y)$ of each modality is independent. By utilizing alternating learning, we can successfully address gradient conflicts in the tabular and image encoders. Since interaction layer is shared, gradient conflicts may arise in shared head during alternating learning, overshadowing the previous modality’s optimization. Hence, to address this issue, we propose a cross-modal gradient surgery that adjusts the current gradients based on cross-modal gradients.

Concretely, assuming the gradient of the current modality at step t is denoted as g , and the gradient from the previous modality is g_p . We define α as the angle between g and g_p . We consider g to be detrimental to the previous modality when

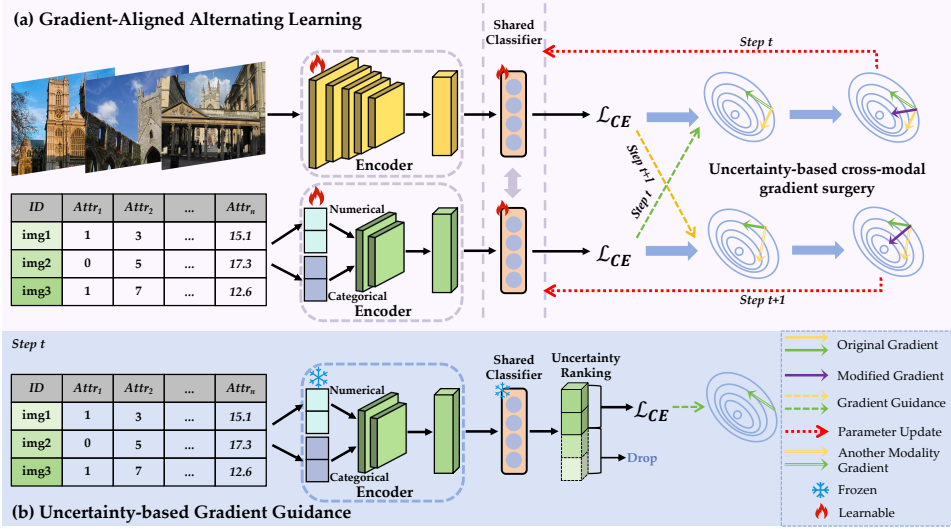


Fig. 2. The framework of the GAAL method. (a) Gradient-Aligned Alternating Learning that only one modality’s learner is updated at each step. Uncertainty-based cross-modal gradient surgery utilizes gradients from cross-modal hard samples to guide the optimization of the shared classifier for the current modality. (b) Uncertainty-based Gradient Guidance that samples hard examples from another modality to provide cross-modal gradient guidance.

$\cos \alpha < 0$. The cosine similarity $\mathcal{S}(\cdot)$ between two gradients g and g_p is

$$\mathcal{S}(g, g_p) = \frac{g^\top g_p}{\|g\|_2 \|g_p\|_2}, \quad (2)$$

where $\|\cdot\|_2$ denote the L_2 norm of vectors.

We aim to project g with the constraint to obtain the modified gradient \tilde{g} that is aligned with g_p :

$$g_p^\top \tilde{g} \geq \epsilon, \quad (3)$$

where ϵ is a small constant that governs the alignment intensity. Inspired by [17], [25], we can formulate (3) as the following optimization problem:

$$\min_{\tilde{g}} \frac{1}{2} \|\tilde{g} - g\|_2^2 \quad (4)$$

$$s.t. \quad g_p^\top \tilde{g} \geq \epsilon.$$

Since the objective function is a convex quadratic function and the constraint is a linear inequality, the problem reduces to a convex quadratic programming problem with a single constraint. Geometrically, \tilde{g} is the projection of g onto g_p , and \tilde{g} is the closest to g in Euclidean distance.

We introduce a dual variable v to represent the gradient weight. By constructing the Lagrangian function and applying KKT conditions, we derive both the closed-form optimal solution as follows:

$$\tilde{g} = g + v g_p, \quad (5)$$

$$v = \max\left(0, \frac{\epsilon - g_p^\top g}{\|g_p\|_2^2}\right).$$

Unlike forcibly orthogonalizing all modal gradients, our method targets only conflicting gradients without enforcing orthogonality, which promotes multimodal synergy. More details are provided in supplemental material.

During alternating iterations, we use cross-modal gradients to guide current modality, which addresses gradient conflicts in the interaction layer and improves cross-modal interaction.

C. Uncertainty-based Gradient Guidance

Through the above gradient surgery, we successfully address gradient conflicts in the interaction layer and improve cross-modal interaction. Subsequently, we aim to select cross-modal gradients to provide more effective guidance in gradient surgery. Since gradients corresponding to hard samples often exhibit larger magnitudes, encapsulate more informative learning signals, and play a more critical role in parameter updates. Therefore, we design Uncertainty-based Gradient Guidance, which selects the gradient direction of high-uncertainty cross-modal samples as the projection direction. This gradient selection not only enables cross-modal gradients to guide current modality but also utilizes current modality to assist in learning cross-modal samples. Learning hard samples effectively enhances the model’s attention to edge cases.

Specifically, for a sample x_i^m in a batch, let p_i^m denote the model’s predicted class probability distribution. We evaluate the uncertainty of each sample by computing the entropy of its predicted probability distribution p_i^m within the batch:

$$\mathcal{H}(p_i^m) = - \sum_{c=1}^Y p_{i,c}^m \log p_{i,c}^m, \quad (6)$$

where $\mathcal{H}(p_i^m)$ is the probability entropy for sample i .

Finally, we select the top of λ^m samples with the highest uncertainty from modality m as hard samples to compute the gradient g_p .

$$\mathcal{I}_p^m = \text{Top}_{\lambda^m}(\mathcal{H}(p_i^m) \mid i \in \mathcal{B}^m), \quad (7)$$

Algorithm 1 GAAL Algorithm.

- 1: **Input:** Training set \mathcal{X} .
 - 2: **Output:** The learned DNN models for all modalities.
INIT Initialize iteration $t = 1$. Initialize encoder parameters θ^I, θ^T and shared classifier parameters Θ .
 - 3: **repeat**
 - 4: **if** $\text{mod}(t, 2) = 1$ **then**
 - 5: Sample $\forall \{x_i^I, x_i^T, y_i\} \in \mathcal{X}$,
 - 6: **end if**
 - 7: Pick up a specific modality $m \in [I, T]$ in order;
 - 8: Calculate predictions p_i^I and p_i^T in (10);
 - 9: Calculate modality m loss \mathcal{L}^m in (1);
 - 10: Update modality-specific encoder parameters θ_i^m ;
 - 11: Calculate modality m classifier gradient g ;
 - 12: Calculate sample entropy $\mathcal{H}^{(I,T)/m}$ in (6);
 - 13: Select the top of $\lambda^{(I,T)/m}$ high-entropy samples;
 - 14: Calculate modality $(I, T)/m$ classifier gradient g_p ;
 - 15: Calculate modified gradient \tilde{g} in (5);
 - 16: Update shared parameters Θ using \tilde{g} ;
 - 17: Update $t = t + 1$;
 - 18: **until** Converge or reach maximum iterations.
-

$$g_p = \frac{1}{|\mathcal{I}_p^m|} \sum_{i \in \mathcal{I}_p^m} \nabla_{\theta} \mathcal{L}(x_i^m), \quad (8)$$

where \mathcal{B}^m denotes the batch of modality m , \mathcal{I}_p^m is the set of indices of the selected top- λ^m high-entropy samples, $\mathcal{L}(x_i^m)$ is the loss for sample i , and ∇_{θ} represents the gradient with respect to classifier parameters. By selecting cross-modal gradients from hard samples, the learning signal can be effectively focused, reducing dilution from easy samples and helping to avoid local optima.

Our algorithm is summarized in Algorithm 1. To sum up, our method GAAL employs alternating learning with shared-head interaction. Subsequently, we aim to address gradient conflicts and enhance synergy in interaction layer by projecting conflicting unimodal gradients. GAAL first selects hard samples from the previous modality to compute cross-modal gradients. Before updating the interaction layer for the current modality, conflicting gradients are projected onto the cross-modal gradients. This promotes hard sample learning and cross-modal interaction while addressing gradient conflicts.

D. Model Inference

During the inference phase, given an tabular-image pair $x_i = [x_i^I, x_i^T]$, we compute the corresponding unimodal logits $f_i^m(x_i^m)$. The unimodal prediction p_i^m is:

$$p_i^m = \text{Softmax}(f_i^m(x_i^m)), \quad (9)$$

and the final multimodal prediction p_i is then derived through a weighted average of these unimodal logits:

$$p_i = \text{Softmax}\left(\frac{1}{2}(f_i^I(x_i^I) + f_i^T(x_i^T))\right). \quad (10)$$

With harmonized tabular-image fusion, inference merely requires simple averaging fusion at the decision level to achieve superior prediction results.

IV. EXPERIMENTS

A. Experimental Settings

Dataset. We conducted experiments on three datasets: Data Visual Marketing (DVM) [8], SUNAttribute [26], and CelebA [27]. **DVM** contains 1,451,784 car images paired with tabular data. Following previous work [23], car models with less than 100 samples were removed, resulting in 286 target classes. **SUNAttribute** is constructed from 717 SUN dataset categories, each annotated with 20 scene attributes. **CelebA** is a facial attribute dataset containing 202,599 face images. More details are provided in supplemental material.

Baselines and Evaluation Metric. We selected various SoTA baselines for comparison, including tabular-image fusion methods: CF [28], MF [29], DAFT [4], TIP [5]; test-time tabular missing methods: KD [30], MFH [31], FMR [32], MMCL [23], CHARMS [24]; multimodal gradient conflict methods: OGM [12], MMPareto [11], MLA [13], DI-MML [15], ReconBoost [14] and LFM [33]. Following the setting of CHARMS [24], we adopt accuracy as the evaluation metric.

Implementation Details. Following MMCL [23], we employ a ResNet50 [34] as image encoder and MLP as the tabular encoder. Additionally, we perform a grid search for hyperparameters and employ early stopping to select the best model. All experiments are conducted on an NVIDIA RTX A6000. More details are provided in supplemental material.

B. Experimental Results

To demonstrate the effectiveness of GAAL, we compare it with multiple popular methods on the three datasets shown in Table I, with results reported as mean values. The “-” indicates that the corresponding method failed to produce results for the modality. In particular, the MFH [31] method fails to handle the complex multi-class classification tasks of the DVM dataset. In summary, from Table I, we can observe that: 1) Compared with unimodal learning, tabular-image fusion methods, multimodal gradient conflict methods and test-time tabular missing methods, GAAL can achieve better performance in almost all cases. 2) GAAL can outperform existing SoTA baselines, achieving not only the best multimodal performance but also the highest unimodal accuracy. 3) The accuracy of test-time tabular missing setting demonstrates that our method achieves the best performance and remains robust. More results are provided in supplemental material.

C. Sensitivity to Hyper-Parameters

We study the impact of the threshold $\{\lambda^I, \lambda^T\}$ on GAAL performance, as shown in Figure 3. An appropriate $\{\lambda^I, \lambda^T\}$ can effectively concentrate gradient signals, providing efficient cross-modal guidance for model optimization. Excessively large or small λ^I, λ^T can have negative effects, either diluting gradient signals or ignoring data distribution.

D. Ablation Study

We perform ablation studies on the DVM and SUNAttribute datasets to analyze the impact of alternating learning, cross-modal gradient surgery (CGS), and uncertainty-based gradi-

TABLE I

THE ACCURACY RESULTS ON DVM, SUNATTRIBUTE, AND CELEBA DATASETS. THE BEST PERFORMANCE IS BOLDDED, AND THE SECOND-BEST IS UNDERLINED. THE IMAGE PERFORMANCE OF TEST-TIME TABULAR MISSING METHODS REPRESENTS THEIR MULTIMODAL PERFORMANCE. THE “*” INDICATES THE VERSION THAT FOLLOWS THE SAME PRETRAINED WEIGHT SETTING AS TEST-TIME TABULAR MISSING METHODS.

Method	DVM			SUNAttribute			CelebA		
	Multi	Image	Tabular	Multi	Image	Tabular	Multi	Image	Tabular
<i>Unimodal Learning</i>									
Resnet50	-	<u>0.8743</u>	-	-	<u>0.8361</u>	-	-	<u>0.8146</u>	-
MLP	-	-	0.8742	-	-	0.8082	-	-	0.7775
<i>Tabular-image Fusion</i>									
CF	0.8793	0.0058	0.8792	0.8278	0.5544	0.8298	0.7740	0.5654	0.7723
MF	0.8993	-	-	0.8333	-	-	0.7941	-	-
DAFT	0.9460	-	-	0.8456	-	-	0.8202	-	-
TIP	0.9545	-	-	<u>0.8612</u>	-	-	<u>0.8224</u>	-	-
<i>Multimodal Gradient Conflict</i>									
OGM	0.8778	0.0076	0.8780	0.8305	0.6457	<u>0.8312</u>	0.7771	0.5328	0.7756
MMPareto	0.9658	0.7549	<u>0.8851</u>	0.8475	0.7992	0.8243	0.8082	0.7941	0.7702
MLA	0.9668	0.7664	0.8505	0.8417	0.8047	0.8194	0.8134	0.7915	0.7913
DI-MML	0.9719	0.8499	0.8748	0.8475	0.8003	0.8212	0.8199	0.8130	0.7839
ReconBoost	0.9714	0.8499	0.8756	0.8498	0.8117	0.8224	0.8149	0.8133	0.7836
LFM	<u>0.9731</u>	0.8543	0.8848	0.8487	0.8075	0.8222	0.8115	0.8110	0.7613
GAAL	0.9917	0.9057	0.9191	0.8668	0.8452	0.8368	0.8273	0.8222	0.7922
<i>Test-time Tabular Missing</i>									
KD	-	0.8390	-	-	0.8382	-	-	0.8118	-
MFH	-	-	-	-	0.8312	-	-	0.7507	-
FMR	-	0.8427	-	-	0.8347	-	-	0.8003	-
MMCL	-	0.8203	-	-	0.8431	-	-	0.8041	-
CHARMS	-	<u>0.9175</u>	-	-	<u>0.8661</u>	-	-	<u>0.8220</u>	-
GAAL*	-	0.9358	-	-	0.8662	-	-	0.8313	-

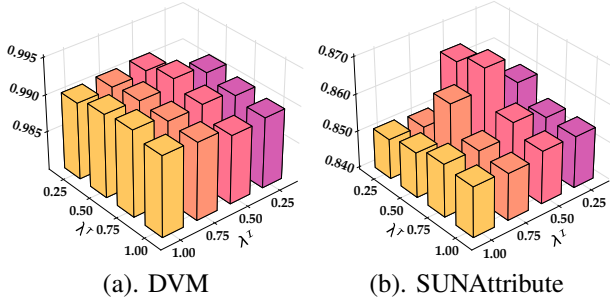


Fig. 3. Sensitivity to hyper-parameter λ^I and λ^T on DVM and SUNAttribute datasets.

TABLE II
RESULTS OF ABLATION STUDIES ON DVM AND SUNATTRIBUTE DATASETS.

CGS	UGG	DVM			SUNAttribute		
		Multi	Image	Tabular	Multi	Image	Tabular
X	X	0.9701	0.8552	0.8760	0.8375	0.7789	0.8103
✓	X	0.9909	0.8931	0.9203	0.8536	0.8201	0.8459
✓	✓	0.9917	0.9057	<u>0.9191</u>	0.8668	0.8452	0.8368

ent guidance (UGG), demonstrating the effectiveness of our method. Experimental results are reported in Table II. From Table II, we can find that: 1) Alternating learning, cross-modal gradient surgery and uncertainty-based gradient guidance can

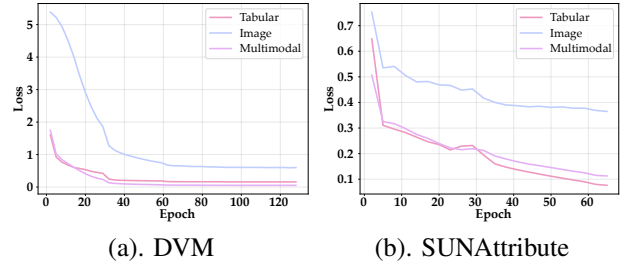


Fig. 4. Convergence results of GAAL on DVM and SUNAttribute datasets.

boost multimodal performance in terms of accuracy. 2) While the unimodal performance of the method using all objectives may not always reach the highest level, it achieves a more balanced classification performance across modalities.

E. Convergence

We also present the convergence results of the model during training on the DVM and SUNAttribute datasets. The loss curves are shown in Figure 4. Since gradients from different modalities naturally decouple, the loss curves of unimodal consistently decrease without getting stuck.

F. Impact of constraint margin ϵ

We vary ϵ on the SUNAttribute dataset and summarize the results in Figure 5. When ϵ is non-negative, it encourages

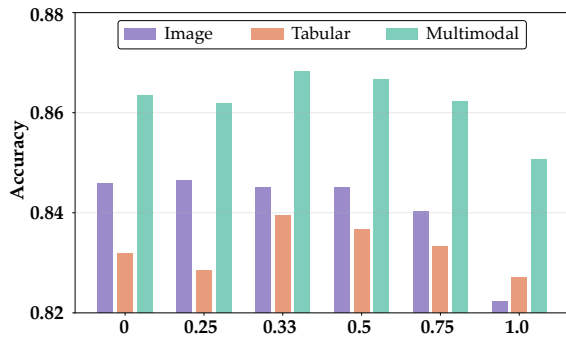


Fig. 5. Impact of constraint margin ϵ .

modality gradient alignment. In particular, setting $\epsilon = 0$ may induce gradient orthogonality, leading to independent updates that are suboptimal for cross-modal interaction. Our results show that properly adjusting ϵ improves performance.

V. CONCLUSION

In this work, we propose GAAL, a novel tabular-image fusion paradigm designed to address multimodal gradient conflicts and facilitate interaction. Our approach adopts alternating learning with a shared classifier to decouple multimodal gradient. To further address gradient conflicts in interaction layer and improve interaction, we design uncertainty-based cross-modal gradient surgery, which utilizes cross-modal gradients to guide the optimization of current modality. Experimental results demonstrate that GAAL outperforms all methods on both image-tabular fusion and test-time tabular missing tasks. We hope this work motivates future research on multimodal challenges encountered in real-world scenarios, with a particular focus on tabular-image learning.

Limitations: Our method focuses on tabular and image classification. Future work will extend the approach to regression and detection tasks.

ACKNOWLEDGMENT

This work was supported in part by the NSFC (62276131), and in part by the Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081).

REFERENCES

- [1] P. Baltescu, H. Chen, N. Pancha, A. Zhai, J. Leskovec, and C. Rosenberg, "Itemsage: Learning product embeddings for shopping recommendations at pinterest," in *KDD*. ACM, 2022, pp. 2703–2711.
- [2] C. Zhang, X. Chu, L. Ma, Y. Zhu, Y. Wang, J. Wang, and J. Zhao, "M3care: Learning with missing modalities in multimodal healthcare data," in *KDD*. ACM, 2022, pp. 2418–2428.
- [3] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical ai," *Nature medicine*, vol. 28, no. 9, pp. 1773–1784, 2022.
- [4] T. N. Wolf, S. Pölsterl, C. Wachinger, A. D. N. Initiative, and A. I. B. L. flagship study of ageing, "DAFT: A universal module to interweave tabular data and 3d images in cnns," *NeuroImage*, vol. 260, p. 119505, 2022.
- [5] S. Du, S. Zheng, Y. Wang, W. Bai, D. P. O'Regan, and C. Qin, "TIP: tabular-image pre-training for multimodal classification with incomplete data," in *ECCV*. Springer, 2024, pp. 478–496.
- [6] H. Pan and Y. Yang, "Coordinated uni-modal assistance for enhancing multi-modal learning," in *ICME*. IEEE, 2025, pp. 1–6.

- [7] Q. Jiang, L. Huang, and Y. Yang, "Rethinking multimodal learning from the perspective of mitigating classification ability disproportion," in *NeurIPS*, 2025.
- [8] J. Huang, B. Chen, L. Luo, S. Yue, and I. Ounis, "DVM-CAR: A large-scale automotive dataset for visual marketing research and applications," in *Big Data*. IEEE, 2022, pp. 4140–4147.
- [9] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *CVPR*. IEEE, 2020, pp. 12 692–12 702.
- [10] Y. Yang, H. Pan, Q. Jiang, Y. Xu, and J. Tang, "Learning to rebalance multi-modal optimization by adaptively masking subnetworks," *TPAMI*, vol. 47, no. 6, pp. 4553–4566, 2025.
- [11] Y. Wei and D. Hu, "Mmpareto: Boosting multimodal learning with innocent unimodal assistance," in *ICML*. PMLR, 2024.
- [12] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *CVPR*. IEEE, 2022, pp. 8228–8237.
- [13] X. Zhang, J. Yoon, M. Bansal, and H. Yao, "Multimodal representation learning by alternating unimodal adaptation," in *CVPR*. IEEE, 2024, pp. 27 456–27 466.
- [14] C. Hua, Q. Xu, S. Bao, Z. Yang, and Q. Huang, "Reconboost: Boosting can achieve modality reconciliation," in *ICML*. PMLR, 2024.
- [15] Y. Fan, W. Xu, H. Wang, J. Liu, and S. Guo, "Detached and interactive multimodal learning," in *ACM MM*. ACM, 2024.
- [16] M. Frank, P. Wolfe *et al.*, "An algorithm for quadratic programming," *Naval research logistics quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.
- [17] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *NeurIPS*, 2017, pp. 6467–6476.
- [18] Y. Gorishniy, I. Rubachev, V. Khurlov, and A. Babenko, "Revisiting deep learning models for tabular data," in *NeurIPS*, 2021, pp. 18 932–18 943.
- [19] A. Margeloiu, X. Jiang, N. Simidjievski, and M. Jamnik, "Tabebm: A tabular data augmentation method with distinct class-specific energy-based models," in *NeurIPS*, 2024.
- [20] V. Borisov, T. Leemann, K. Sebler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *TNNLS*, vol. 35, no. 6, pp. 7499–7519, 2024.
- [21] Q. Jiang, Z. Chi, and Y. Yang, "Interactive multimodal learning via flat gradient modification," in *IJCAI*. ijcai.org, 2025, pp. 5489–5497.
- [22] J. Wu and M. Harandi, "Munba: Machine unlearning via nash bargaining," in *ICCV*. IEEE, 2025, pp. 4754–4765.
- [23] P. Hager, M. J. Menten, and D. Rueckert, "Best of both worlds: Multimodal contrastive learning with tabular and imaging data," in *CVPR*. IEEE, 2023, pp. 23 924–23 935.
- [24] J. Jiang, H. Ye, L. Wang, Y. Yang, Y. Jiang, and D. Zhan, "Tabular insights, visual impacts: Transferring expertise from tables to images," in *ICML*. PMLR, 2024.
- [25] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with A-GEM," in *ICLR*. OpenReview, 2019.
- [26] G. Patterson, C. Xu, H. Su, and J. Hays, "The SUN attribute database: Beyond categories for deeper scene understanding," *IJCV*, vol. 108, no. 1-2, pp. 59–81, 2014.
- [27] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*. IEEE, 2015, pp. 3730–3738.
- [28] S. Spasov, L. Passamonti, A. Duggento, P. Lio, N. Toschi, A. D. N. Initiative *et al.*, "A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease," *Neuroimage*, vol. 189, pp. 276–287, 2019.
- [29] L. A. Vale-Silva and K. Rohr, "Long-term cancer survival prediction using multimodal deep learning," *Scientific Reports*, vol. 11, no. 1, p. 13505, 2021.
- [30] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.
- [31] Z. Xue, Z. Gao, S. Ren, and H. Zhao, "The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation," in *ICLR*. OpenReview, 2023.
- [32] Y. Yang, D. Zhan, Y. Fan, Y. Jiang, and Z. Zhou, "Deep learning for fixed model reuse," in *Proceedings of the AAAI Conference on Artificial Intelligence*, S. Singh and S. Markovitch, Eds. AAAI Press, 2017, pp. 2831–2837.
- [33] Y. Yang, F. Wan, Q. Jiang, and Y. Xu, "Facilitating multimodal classification via dynamically learning modality gap," in *NeurIPS*, 2024.
- [34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*. IEEE, 2016, pp. 770–778.