Multimodal Inference with Incremental Tabular Attributes

Xinda Chen , **Zhen Xing** , **Zixian Zhang** , **Weimin Tan**[*] and **Bo Yan**[*]
Shanghai Key Laboratory of Intelligent Information Processing,
Computation and Artificial Intelligence Innovative College, Fudan University, Shanghai, China
{chenxd23, zhangzx23}@m.fudan.edu.cn, {xingz20, wmtan, byan}@fudan.edu.cn,

Abstract

Multimodal Learning with visual and tabular modalities has become more and more popular nowadays, especially in the healthcare area. Due to the adaptation of new equipment or new factors being introduced, the tabular modality keeps changing. However, the standard process of training multimodal AI models requires tables to have fixed columns in training and inference; thus, it is not suitable for handling dynamically changed tables. Therefore, new methods are needed for efficiently handling such tables in multimodal learning. In this paper, we introduce a new task, multimodal inference with incremental tabular attributes, which aims to enable trained multimodal models to leverage incremental attributes in tabular modality during the inference stage efficiently. We implement a specialized encoder to disentangle the latent representation of incremental tabular attributes inside itself and with the old attributes to reduce information redundancy and further align the incremental attributes with the visual modality with consistency loss to improve information richness. Experimental results across five public datasets show that our method effectively utilizes incremental tabular attributes, achieving state-of-the-art performance in general scenarios. Beyond the inference, we also find that our method achieved better performance in fully supervised settings, evoking a new training style for multimodal learning with tables.

1 Introduction

Modern datasets often include multiple modalities, with visual and tabular modalities being particularly prevalent in the medical field [Sudlow *et al.*, 2015]. The visual modality is typically derived from imaging techniques such as CT scans, containing rich spatial and anatomical information. In contrast, the tabular modality often records structured patient data, such as medical history, laboratory results, and demographic information, which encapsulates a different set of critical features compared to images [Soenksen *et al.*, 2022; Yi *et al.*, 2023]. Numerous studies have demonstrated the efficacy of integrating visual and tabular modalities, showing that such multimodal approaches often outperform single-modality models [Huang, 2023; Hager *et al.*, 2023]. However, existing methods generally assume a static tabular modality, requiring the attributes of tabular data at inference time to align with those present during training [Somepalli *et al.*, 2021]. In practical applications, tabular data is frequently updated. For instance, in Alzheimer's disease prediction, existing models have been trained on MRI images with demographic information and clinical indicators. Recently, significant biomarkers such as the YWHAG protein [Guo *et al.*, 2024] and inflammation/immune-related markers [Jack *et al.*, 2024] have been discovered. Leveraging these new biomarkers alongside existing models for retraining poses a challenge, as recalling patients from previous databases for re-evaluation is impractical. Such dynamic nature of tabular modalities underscores the need for models capable of handling incremental tabular attributes in a multimodal context, enabling more robust and adaptable inference in real-world scenarios.

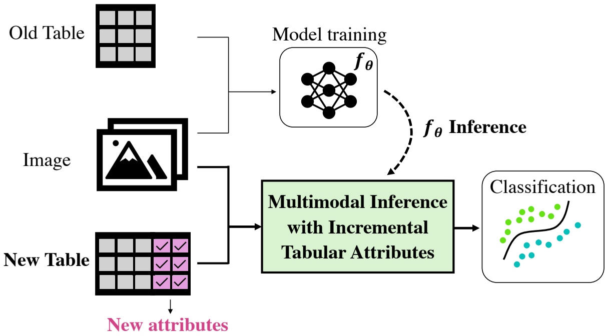


Figure 1: Illustration of multimodal inference with incremental tabular attributes. A trained model on old table and image modalities encounters incremental attributes (e.g. new biomarkers [Jack *et al.*, 2024; Guo *et al.*, 2024] highlighted in pink) in tabular inputs during inference. The goal is to seamlessly incorporate these incremental attributes into the model to adapt to dynamic real-world tabular environments, overcoming limitations of existing methods.

To address this challenge, we introduce a novel task termed multimodal inference with incremental tabular attributes. As shown in Figure 1, we consider a scenario in which a multi-

[*] Weimin Tan and Bo Yan are the corresponding authors.

modal model has been trained using any deep learning algorithm on visual and tabular modalities. During the inference stage, new attribute columns are introduced to the tabular modality. Our objective is to propose a method that fine-tunes the trained model to leverage the relationships among the incremental tabular attributes, the old tabular attributes, and the visual modality. In doing so, the model learns task-relevant information more effectively in an unsupervised manner. This approach aims to outperform models that discard the new attributes and rely solely on the old ones, while simultaneously reducing the data preparation overhead and improving convergence speed compared to fully supervised re-training.

Designing such a method is challenging, with the main difficulty being the inability to dynamically adapt due to the heterogeneity of the tabular modality. Each column in a table represents a distinct attribute with varying scales, semantics, and statistical distributions [Nam *et al.*, 2023]. Directly applying encoding patterns from old attributes to new ones may not only fail to improve performance but could also degrade it. Furthermore, effectively leveraging the visual and tabular modalities to compress redundant information in incremental attributes and extract novel, task-relevant knowledge is essential. This ensures the new attributes contribute meaningfully to the model. However, research on multimodal tasks involving visual and tabular modalities, particularly when incremental tabular attributes are introduced, remains limited.

To address the aforementioned challenges, we propose a novel framework named Multimodal Inference with Incremental Tabular Attributes (MIITA). Specifically, we design a dedicated encoder for the incremental tabular attributes based on a Variational Autoencoder (VAE). This encoder not only extracts implicit features but also decouples them, ensuring that the learned latent representations are as independent as possible across dimensions. This disentanglement enhances interpretability and makes the latent features more amenable to constraints imposed by the old tabular attributes and the visual modality. Besides classification loss and CLIP loss—encouraging the representations of the two modalities to be close for the same sample and distant for different—we introduce two new loss functions. The first is a disentangled loss designed to further improve the disentanglement of the latent representations with the old representation. The second is a consistency loss, aimed at strengthening the alignment between the incremental tabular attributes and the visual modality. Both losses have demonstrated their effectiveness in ablation, showing significant contributions to overall performance. In summary, the contributions of this paper are:

- This paper introduces a new task, Multimodal Inference with Incremental Tabular Attributes, aiming to enable multimodal models to incorporate incremental tabular columns during the inference stage, enhancing the practicality of AI models in dynamically changed tables.
- To dynamically and unsupervisedly leverage incremental tabular attributes, we propose the MIITA framework, which disentangles the incremental representation from the old ones and adds the modality consistency constraint for better alignment. Our approach achieves state-of-the-art performance across five public datasets.
- Despite the new task, we show that MIITA evokes a new training procedure for general supervised multimodal learning with tabular and vision modalities.

2 Related Work

2.1 Multimodal Contrastive Learning

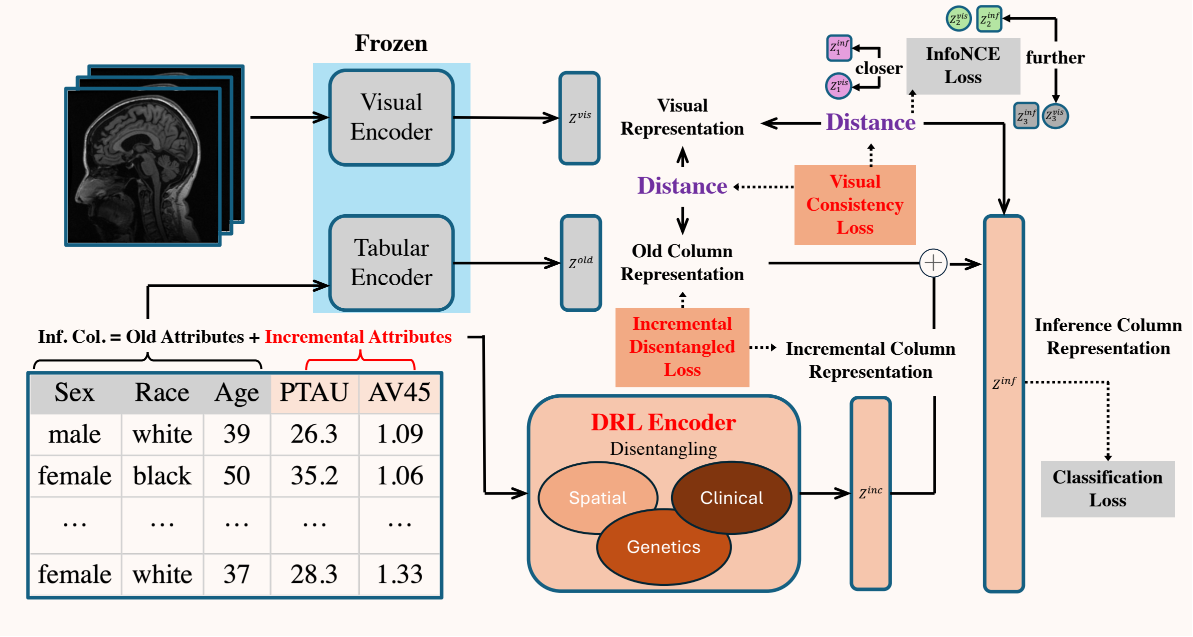
Multimodal contrastive learning has emerged as a powerful framework for learning representations across different modalities. Methods such as CLIP [Radford *et al.*, 2021] and ALIGN [Jia *et al.*, 2021] leverage large-scale paired datasets to align representations of text and images through contrastive objectives. These approaches ensure that similar samples across modalities are brought closer in the shared representation space, while dissimilar samples are pushed apart [Wu *et al.*, 2021]. Recent works have extended this paradigm to other modality pairs, especially visual-tabular data [Hager *et al.*, 2023; Wang *et al.*, 2024a], demonstrating its versatility in fusing visual and tabular information.

However, existing multimodal contrastive learning methods are generally designed for static multimodal datasets and assume fixed input dimensions for each modality. When new attributes are introduced to the tabular modality, these models fail to adapt without retraining on the updated dataset. This limitation hinders their applicability to real-world scenarios.

2.2 Tabular Deep Learning

Tabular deep learning has become an important area of research, focusing on adapting neural networks to structured tabular data. Architectures such as TabNet [Arik and Pfister, 2020], FT-Transformer [Gorishniy *et al.*, 2023] and SCARF [Bahri *et al.*, 2022] have shown promise in feature selection, representation learning, and interpretability for tabular data. These models leverage advanced attention mechanisms or feature gating to capture meaningful relationships between attributes and outperform traditional machine learning approaches in various tasks. Recently, Large Language Models (LLMs) have also been explored for tabular data tasks. Methods like TabLLM [Hegselmann *et al.*, 2023a] and TableGPT2 [Su *et al.*, 2024] convert tabular attributes into natural language formats, enabling LLMs to incorporate prior knowledge and perform reasoning directly on structured data.

Despite these advancements, traditional tabular deep learning models typically assume a fixed tabular input. Current efforts, such as TransTab [Wang and Sun, 2022] and TabPFN [Hollmann *et al.*, 2025], have introduced methods for handling dynamic attributes in tabular data. However, these approaches are predominantly focused on single-modality settings and lack the capability to effectively utilize visual modality information. On the other hand, LLM-assisted approaches excel in zero-shot or few-shot scenarios by leveraging pre-trained knowledge but often fall short in scenarios requiring incremental learning of incremental attributes [Zhang *et al.*, 2024; Carballo *et al.*, 2023]. Consequently, both traditional and LLM-assisted methods face significant challenges in addressing dynamic tabular attributes and fully leveraging cross-modal relationships, leaving a gap in solving incremental multimodal inference tasks.



Sex	Race	Age	PTAU	AV45
male	white	39	26.3	1.09
female	black	50	35.2	1.06
...	...	...	...	...
female	white	37	28.3	1.33

Figure 2: Overview of MIITA framework. MIITA takes images and an inference table with incremental attributes as input, producing adapted inference column representations for downstream tasks. A trained multimodal model provides frozen tabular and visual encoders to extract the visual and old tabular representations. A novel Disentangled Representation Learning (DRL) encoder processes incremental tabular attributes, disentangling them into interpretable dimensions. The framework leverages four loss components: InfoNCE Loss, Classification Loss with pseudo-labels, Incremental Disentangled Loss, and Visual Consistency Loss for modality alignment. This design enables efficient use of incremental tabular attributes, improving performance and convergence in an unsupervised manner.

3 Method

In this section, we formally define the task of Multimodal Inference with Incremental Tabular Attributes (MIITA). We then present the proposed framework, also named MIITA, detailing its architecture and the mechanisms it employs to handle incremental updates in tabular attributes during inference.

3.1 Problem Formulation

Consider a multimodal dataset $\mathcal{D} = \{(x_i^{\text{vis}}, x_i^{\text{old}}, y_i)\}_{i=1}^N$, where $x_i^{\text{vis}} \in \mathcal{X}^{\text{vis}}$ represents a visual input (e.g., medical images), $x_i^{\text{old}} \in \mathcal{X}^{\text{old}}$ denotes tabular data comprising M old attributes, and $y_i \in \mathcal{Y}$ is the target label. A multimodal model f is trained on $\mathcal{D}$ to predict y_i by jointly utilizing x_i^{vis} and x_i^{old}:

$$\hat{y}_i = f(x_i^{\text{vis}}, x_i^{\text{old}}; \Theta), \tag{1}$$

where Θ denotes the model parameters learned in training.

In the MIITA setting, during the inference phase, incremental attributes are added to the tabular modality, resulting in an incremental tabular input $x_i^{\text{inf}} \in \mathcal{X}^{\text{inf}}$, where $\mathcal{X}^{\text{inf}} \supset \mathcal{X}^{\text{old}}$ and the incremental attributes are denoted as $\mathcal{X}^{\text{inc}} = \mathcal{X}^{\text{inf}} \setminus \mathcal{X}^{\text{old}}$. The goal is to adapt the trained model f to utilize the augmented tabular input x_i^{inf} along with x_i^{vis} to predict y_i:

$$\hat{y}_i = f'(x_i^{\text{vis}}, x_i^{\text{inf}}; \Theta'), \tag{2}$$

where f' and Θ' represent the adapted model and updated parameters, respectively.

The MIITA task involves three key challenges:

- **Dynamic Adaptation**: The model must incorporate $\mathcal{X}^{\text{inc}}$ without requiring retraining on the entire dataset.
- **Preserving Knowledge**: The model should retain its ability to utilize the old attributes $\mathcal{X}^{\text{old}}$ and $\mathcal{X}^{\text{vis}}$.
- **Unsupervised Adaptation**: The adaptation should not rely on labels for the incremental attributes, instead leveraging unsupervised learning to extract task-relevant information.

By addressing these challenges, MIITA enables robust multimodal inference in dynamic real-world scenarios where tabular data evolves over time.

3.2 MIITA Framework

As shown in Figure 2, our framework is based on a trained multimodal model, from which we extract the tabular encoder and the visual encoder, freezing their parameters to ensure that MIITA does not perform worse than the old model. During inference, the image and the old tabular attributes of a given sample are passed through the visual encoder and tabular encoder, respectively, producing the visual representation and old tabular representation.

To handle incremental tabular attributes, we use a dedicated **Disentangled Representation Learning (DRL)** encoder. This encoder disentangles the entangled representations of the incremental attributes into interpretable dimensions, such as spatial or temporal factors. This disentanglement not only enhances the interpretability of the model but also strengthens the effectiveness of subsequent multimodal learning procedures. After passing through the DRL encoder, we obtain the incremental tabular representation.

Our framework leverages the relationships among the three representations: visual, old tabular representation, and incremental tabular representation, using a combination of novel loss functions. Beyond the foundational InfoNCE loss and a task-specific classification loss with pseudo-labels, we will introduce novel Incremental Disentangled Loss and Visual Consistency Loss in detail. By integrating the DRL encoder with these four loss functions, we finally use the improved inference tabular representation for downstream tasks. Our framework can effectively utilize the information from the incremental tabular attributes, improving model performance in an unsupervised manner while achieving faster convergence.

DRL Encoder

The DRL encoder is pivotal for disentangling incremental tabular attributes, enabling effective utilization of new information in the MIITA framework. Specifically, it employs a β-Variational Autoencoder (β-VAE) to encode the incremental attributes x^{inc} into a disentangled latent space z^{inc}, ensuring independence among dimensions [Higgins *et al.*, 2017]. Other DRL encoder variants, like FactorVAE [Kim and Mnih, 2019] and DIP-VAE [Kumar *et al.*, 2018], are also compatible with our framework. The β-VAE's objective is defined by a combination of reconstruction and regularization losses.

The reconstruction loss ensures that the encoder effectively captures the information contained in x^{inc}:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{q_\phi(z^{\text{inc}}|x^{\text{inc}})}\left[\|x^{\text{inc}} - \hat{x}^{\text{inc}}\|^2\right], \tag{3}$$

where $\hat{x}^{\text{inc}}$ is the reconstruction of the input. To promote disentanglement, a regularization term enforces the independence of latent dimensions by introducing a KL divergence between the posterior $q_\phi(z^{\text{inc}}|x^{\text{inc}})$ and a factorized unit Gaussian prior $p(z^{\text{inc}})$:

$$\mathcal{L}_{\text{KL-inc}} = D_{\text{KL}}(q_\phi(z^{\text{inc}}|x^{\text{inc}})\|p(z^{\text{inc}})). \tag{4}$$

The overall loss for the DRL encoder is expressed as:

$$\mathcal{L}_{\text{DRL}} = \mathcal{L}_{\text{rec}} + \beta\mathcal{L}_{\text{KL-inc}}, \tag{5}$$

where β is a hyperparameter that controls the trade-off between reconstruction accuracy and disentanglement strength.

Incremental Disentangled Loss

In addition to disentangling the incremental attributes themselves, we aim to ensure that z^{inc} encodes meaningful, complementary information while avoiding redundancy with z^{old}. To achieve this, we introduce a disentangled loss with two distinct focuses:

The **redundancy mitigation loss** focuses on maximizing the KL divergence between $q_\phi(z^{\text{inc}}|x^{\text{inc}})$ and $q(z^{\text{old}})$, ensuring z^{inc} captures novel, nonlinear information not present in z^{old}:

$$\mathcal{L}_{\text{KL-redundancy}} = -D_{\text{KL}}(q_\phi(z^{\text{inc}}|x^{\text{inc}})\|q(z^{\text{old}})). \tag{6}$$

The **covariance regularization loss**, on the other hand, penalizes linear cross-correlations between the dimensions of z^{inc} and z^{old}, further promoting disentanglement by encouraging independence:

$$\mathcal{L}_{\text{cov-cross}} = \sum_{i,j}\left(C_{ij}^{\text{cross}}\right)^2, \tag{7}$$

where C_{ij}^{cross} is the cross-covariance matrix in z^{inc} and z^{old}.

The overall disentangled loss is then defined as:

$$\mathcal{L}_{\text{Disentangled}} = \lambda_1\mathcal{L}_{\text{KL-redundancy}} + \lambda_2\mathcal{L}_{\text{cov-cross}}, \tag{8}$$

where λ_1 and λ_2 are hyperparameters controlling the trade-off between nonlinear redundancy mitigation and linear covariance regularization.

InfoNCE Loss

The InfoNCE loss [van den Oord *et al.*, 2019] facilitates alignment between the visual and tabular modalities by encouraging the inference tabular representation z^{inf}, formed by concatenating z^{old} and z^{inc}, to be close to the visual representation z^{vis} of the same sample while being far from those of other samples in the batch.

The inference tabular representation is defined as:

$$z^{\text{inf}} = \text{concat}(z^{\text{old}}, z^{\text{inc}}). \tag{9}$$

The InfoNCE loss is expressed as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\exp\left(\text{sim}\left(z_b^{\text{inf}}, z_b^{\text{vis}}\right)/\tau\right)}{\sum_{k=1}^{B}\exp\left(\text{sim}\left(z_b^{\text{inf}}, z_k^{\text{vis}}\right)/\tau\right)}, \tag{10}$$

where $\text{sim}(a, b)$ denotes cosine similarity, τ is a temperature hyperparameter, and B is the batch size. This loss ensures that z^{inf} aligns closely with z^{vis} for the same sample while preserving inter-sample separability.

Visual Consistency Loss

To better align the tabular and visual modalities, we further ensure that the inference tabular representation z^{inf} is closer to the visual representation z^{vis} than the old tabular representation z^{old} alone. This encourages incremental tabular attributes to complement the old tabular representation effectively.

To enforce the desired alignment, we propose an InfoNCE-inspired loss that compares the relative similarities between the representations:

$$\mathcal{L}_{\text{Vis-Cons}} = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{N}{N+D}, \tag{11}$$

$$N = \exp\left(\text{sim}\left(z_b^{\text{inf}}, z_b^{\text{vis}}\right)/\tau\right), D = \exp\left(\text{sim}\left(z_b^{\text{old}}, z_b^{\text{vis}}\right)/\tau\right).$$

By minimizing $\mathcal{L}_{\text{Vis-Cons}}$, the model explicitly enforces the inference tabular representation z^{inf} to capture richer and more complementary information from the incremental attributes, improving its alignment with the visual modality. This approach leverages the additional information from incremental tabular attributes to enhance multimodal inference while maintaining consistency with the visual modality.

Classification with Pseudo Labels

To enhance the framework's classification capabilities, pseudo-labels generated by old models are employed to provide supervision in an unsupervised setting. To ensure reliability, only pseudo-labels with a confidence score higher than a specific threshold (e.g., 0.7) are retained for supervision. The classification loss is defined using the standard cross-entropy loss as [Mao *et al.*, 2023]:

$$\mathcal{L}_{\text{Cls}} = -\frac{1}{B}\sum_{i=1}^{B}\hat{y}_i\log y_i, \tag{12}$$

where $\hat{y}_i$ represents the highest pseudo-label probability of sample i, y_i is the predicted probability for the corresponding class.The purpose of $\mathcal{L}_{\text{Cls}}$ is to ensure that the learned features are strongly related to downstream tasks by aligning them with task-relevant information. This process helps to discard task-irrelevant features, thereby improving the effectiveness of MIITA for specific applications.

Overall Objective

The overall objective integrates all components to optimize the MIITA framework holistically:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{DRL}} + \mathcal{L}_{\text{Disentangled}} + \mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{Vis-Cons}} + \mathcal{L}_{\text{Cls}} \quad (13)$$

where each term contributes to specific aspects of representation learning, alignment, and classification. By balancing these losses, the model effectively captures and integrates multimodal information, achieving robust and interpretable performance.

4 Experiment

4.1 Experiment Setup

Datasets and Comparative Methods

We use five public datasets that are commonly used in multimodal learning areas using tabular and visual modalities: ADNI (AD) [Jr *et al.*, 2008], Data Visual Marketing (DV) [Huang *et al.*, 2023], Pokemon Primary Type (PK), HearthStone Card's category (HS), CS:GO skin quality (CG) [Lu *et al.*, 2023]. In each, we remove specific columns from the train and validation sets while keeping them in the inference set to simulate real-world incremental tabular inference. The deleted columns were chosen logically, reflecting their historical appearance. Details are in the appendix.

For baselines, we use single tabular modality classifiers: XGBoost [Chen and Guestrin, 2016], SCARF [Bahri *et al.*, 2021] and FT-Transformer [Gorishniy *et al.*, 2021]. Since these tabular methods can't utilize incremental columns, we only use the same columns as the train set for inference. Single visual modality classifiers like CNN [O'Shea and Nash, 2015] and ViT [Dosovitskiy *et al.*, 2021] are also used. Then, we compare our method to recent tabular-visual multimodal models: AutoMM [Shi *et al.*, 2021], SimCLR [Chen *et al.*, 2020], MMCL [Huang, 2023] and TIP [Du *et al.*, 2024]. Also, we compare our method to FT-Trans* and TransTab [Wang and Sun, 2022], which can utilize all columns in the inference set. Note that in our experiments, we modified FT-Transformer by replacing the embedding layer with a linear embedding (1 × encoder size) to handle any input dimensions, noted as FT-Trans*. Finally, we compare performance with LLM-assisted models, including TabLLM [Hegselmann *et al.*, 2023b] and MediTab [Wang *et al.*, 2024b].

Implementation Details

We encode categorical features using a backward difference encoder [Potdar *et al.*, 2017]. Key setting of MIITA are: β (VAE) $\in \{2, 5, 10\}$, λ_1 (KL) / λ_2 (cov-cross) $\in \{0.5, 1.0, 2.0\}$, and pseudo-label threshold = 0.7. The loss weights in Eq.13 were adjusted based on early gradient magnitudes. The default setting is $\{1, 0.5, 0.8, 0.4, 1\}$.

	Modality Input		Model	Predicted Result		Target
Old tabular attributes	Year	2018	XGBoost	BMW 1 series	✗	Car Brand
	Color	White				
	...	...				
Image			ViT	Volkswagen Golf	✗	
Incremental tabular attributes	Price	27000	MMCL	Volkswagen Golf	✗	
	Speed	136 mph	MIITA	Audi A3	✓	
	...	...				
Old tabular attributes	HP	100	XGBoost	Dark	✗	Primary Type
	Attack	134				
	...	...				
Image			ViT	Dragon	✗	
Incremental tabular attributes	Weight	202 kg	MMCL	Dark	✗	
	Ability	Sand Stream	MIITA	Rock	✓	
	...	...				

Figure 3: Visualization of the MIITA task. In the car brand and Pokémon primary type prediction tasks, existing methods fail to make accurate classifications due to their inability to effectively utilize incremental tabular attributes. In contrast, our MIITA approach successfully learns from this information, leading to correct results.

4.2 Results on Public Datasets

We evaluated the MIITA framework for six tasks in five widely used datasets that span entirely different domains, each containing both tabular and visual modalities. As shown in Table 1, MIITA consistently outperformed state-of-the-art (SOTA) single-modal and multimodal methods, representing MIITA is a general framework applicable for all scenarios including medical, advertisement and game. The visualization result of two difficult samples is shown in Figure 3. All results were averaged over four runs to mitigate randomness, with the corresponding variances provided in the appendix.

Single-Modal Baselines

We first conducted experiments using single-modal models for both tabular and visual data. Since traditional tabular learning methods cannot handle dynamic attributes, our tabular experiments only used the static, old attributes matching the training setup. The results demonstrate significant performance limitations in both modalities: the visual data proved unstable and informationally limited, while the tabular modality was severely constrained by its inability to utilize incremental columns.

Multimodal Methods

Three multimodal learning models designed for tabular and visual modalities were evaluated. These models achieved superior performance compared to single-modal baselines, highlighting the effectiveness of multimodal learning in leveraging complementary information from both modalities. We used SimCLR as a baseline for further experiments, replacing its tabular feature extractor with methods capable of handling dynamic attributes.

Experiments with Incremental Tabular Attributes

To explore handling dynamic tabular attributes, we first tested FT-Trans*, which directly maps tabular data with incremental attributes using the old encoder without adaptation. The

Dataset size	1250	1250	176414	897	10710	956	
Train tabular attribute size	30	30	5	7	5	2	
Incremental tabular attribute size	85	85	11	10	7	3	
Method/Dataset	**AD3**	**AD2**	**DV**	**PK**	**HS**	**CG**	**Rank(Std)**
Tabular classifiers							
XGBoost	0.758	0.866	0.890	0.616	0.600	0.546	9.8 (1.47)
SCARF	0.742	0.852	0.887	0.589	0.600	0.521	11.2 (1.51)
FT-Trans	0.746	0.868	0.891	0.607	0.610	0.533	9.2 (2.27)
Visual classifiers							
CNN	0.761	0.883	0.880	0.322	0.550	0.682	8.8 (4.54)
ViT	0.768	0.885	0.880	0.308	0.568	0.674	8.6 (4.59)
Multimodal classifiers							
AutoMM	0.760	0.902	0.887	0.620	0.460	0.600	8.0 (2.85)
SimCLR	0.760	0.908	0.896	0.634	0.702	0.604	4.3 (1.86)
MMCL	0.773	0.908	0.898	0.646	0.689	0.610	3.2 (0.98)
Multimodal classifiers with traditional dynamic tabular encoder							
SimCLR w. FT-Trans*	0.628	0.830	0.847	0.561	0.432	0.519	13.7 (0.82)
SimCLR w. TransTab	0.730	0.882	0.892	0.634	0.647	0.590	7.4 (3.32)
TIP	0.788	0.908	0.906	0.634	0.653	0.624	3.2 (0.98)
Multimodal classifiers with LLM-assisted tabular encoder							
SimCLR w. TabLLM	0.742	0.868	0.885	0.617	0.602	0.586	9.7 (1.54)
SimCLR w. MediTab	0.760	0.896	0.892	0.620	0.605	0.580	7.0 (1.58)
MIITA	**0.814**	**0.930**	**0.924**	**0.692**	**0.714**	**0.742**	**1.0 (0.00)**

Table 1: Comparison of different methods on public datasets. AD3 denotes a three-class classification on AD dataset, whereas AD2 refers to a binary classification. The evaluation index is accuracy. For each dataset, the best results are shown in bold. Reported results are averaged over four trials. The rank column reports the average rank across all datasets.

results were poor, demonstrating the necessity for tailored adjustments in encoding schemes. Both the traditional self-supervised pretraining approach (TransTab, TIP) and large language model (LLM)-assisted tabular learning also failed to achieve satisfactory results. Their performance was often worse than multimodal models using only the old tabular attributes because they cannot put trained tabular encoders into fully usage. In contrast, MIITA consistently achieved leading performance across all tasks. This demonstrates the superiority and generalizability of the MIITA framework in addressing the challenges posed by incremental tabular attributes in multimodal learning.

4.3 Generalizability Across Different Scenarios

As discussed in the methodology section, MIITA is designed to be a model-agnostic framework, capable of integrating with various architectures seamlessly. As illustrated in Figure 4, MIITA demonstrates consistent performance improvements when applied to models employing different tabular and visual feature extractors in AD3 dataset. This showcases the flexibility of MIITA in adapting to diverse base architectures without requiring significant structural modifications.

The generalizability of MIITA extends beyond tabular-visual multimodal tasks to scenarios involving other modalities. We conducted experiments on additional multimodal learning tasks: tabular-text in a game dataset[Lu *et al.*, 2023] and tabular-temporal data in a stock dataset. Across these tasks, MIITA achieved a consistent performance improvement, highlighting its effectiveness in leveraging incremental tabular attributes with other modalities, reinforcing its potential as a universal approach for dynamic multimodal learning.

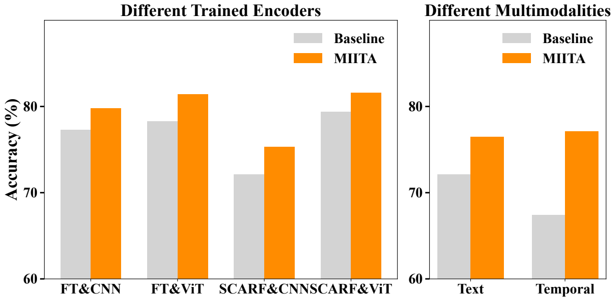


Figure 4: Performance comparison of MIITA across different scenarios. MIITA demonstrates consistent performance improvements when applied to different encoders with FT-Trans, SCARF in Tables and CNN, ViT in Visions as well as different multimodalities, including tabular-text and tabular-temporal, highlighting MIITA's flexibility and generalizability in integrating with diverse base models without requiring significant structural modifications.

4.4 General Improvement on Supervised Learning

Beyond excelling in the proposed MIITA task, our framework also demonstrated significant advantages in fully su-

Settings/Dataset	AD3	PK	DV
Fully supervised	0.838	0.774	0.930
Disentangled rate	0.343	0.457	0.678
MIITA	**0.870**	**0.796**	**0.938**
Disentangled rate	0.689	0.824	0.749

Table 2: Performance comparison between MIITA and existing multimodal learning models under fully supervised scenarios on three datasets. Results demonstrate that the progressive inference strategy using the MIITA framework achieves superior accuracy, leveraging disentangled tabular features for incremental integration.

Attributes/Dataset	AD3	DV	PK
High Visual-Corr	0.783	0.900	0.646
Medium Visual-Corr	0.794	0.900	0.660
Low Visual-Corr	**0.814**	**0.920**	**0.689**
Few	0.788	0.900	0.668
Moderate	**0.820**	0.908	0.683
Many	0.814	**0.924**	**0.692**

Table 3: MIITA performance on various incremental tabular attribute sets, top three rows represent the attributes' correlation with visual modality, while the bottom three rows represent the number of tabular attributes.

pervised multimodal scenarios. The experiments showed that MIITA consistently outperformed traditional training approaches when labeled data was available.

In the supervised setting, the MIITA framework was applied by first disentangling the complete tabular data using the DRL encoder. The disentangled features were then treated as original inputs and progressively grouped in batches to train the MIITA framework with the proposed loss constraints. For this scenario, the DRL encoder utilized a Transformer-based architecture. As presented in Table 2, this incremental training approach achieved superior performance across three datasets compared to state-of-the-art multimodal learning models [Hager *et al.*, 2023]. Furthermore, we evaluated the disentanglement level using the metric introduced in [Higgins *et al.*, 2017], demonstrating that MIITA achieves a higher disentanglement rate compared to traditional multimodal methods, which is beneficial for the integration of multimodal learning. These findings highlight the broader applicability of the MIITA framework and its potential to inspire novel strategies for progressive training and disentangling in general tabular-visual multimodal learning tasks, moving beyond the limitations of holistic learning approaches.

4.5 Incremental Tabular Attributes Settings

As shown in Table 3, we explored the impact of different incremental attribute settings in the tabular modality on the performance of MIITA. Based on human cognitive perception, we categorized the tabular attributes into three groups: those highly related to the image (i.e., the table attributes are fully reflected in the image), moderately related, and weakly related. Experimental results demonstrate that weakly related tabular attributes provide more information for multimodal tasks and play a more significant role in incremental inference. We also tested different numbers of incremental attributes, categorized as few, moderate, and many. Results indicate that as more attributes are added, MIITA's performance generally will improve. Even with a small incremental column size, MIITA shows performance gains, meaning MIITA treats incremental attributes as valuable rather than noise.

Components/Dataset	AD3	DV	PK
w/o. VCL & DRL & IDL	0.788	0.900	0.651
w/o. VCL	0.796	0.906	0.662
w/o. DRL & IDL	0.790	0.906	0.679
w/o. IDL	0.773	0.904	0.608
ALL	**0.814**	**0.924**	**0.692**

Table 4: Ablation study of different components in MIITA, VCL represents Vision Consistency Loss, DRL represents Disentangled Representation Learning Encoder while IDL represents Incremental Disentangled Loss.

4.6 Ablation Studies

To validate the effectiveness of the MIITA framework and the proposed loss functions, we conducted ablation studies by removing or replacing key components, as summarized in Table 4. Removing the visual consistency loss caused a notable performance drop, emphasizing its role in aligning the tabular representation with the visual modality. Replacing the disentangled encoder and incremental disentangled loss with a standard transformer also led to reduced performance, highlighting the benefits of disentangling incremental tabular attributes into interpretable dimensions. Additionally, using the disentangled encoder without the incremental disentangled loss caused a significant performance decline, showing that isolating incremental attributes leads to redundancy and inefficiency. The full MIITA framework, with all modules included, achieved the best results, demonstrating the effectiveness of each component in addressing the challenges of incremental tabular attributes in multimodal learning.

5 Discussion and Conclusion

In this work, we introduced MIITA, a novel framework designed to address the challenge of multimodal inference with incremental tabular attributes. Our extensive experiments demonstrated that MIITA outperforms state-of-the-art single-modal and multimodal methods across a variety of datasets. The key strength of MIITA lies in its ability to effectively integrate incremental tabular attributes, leveraging disentangled representation learning and innovative loss functions to maintain alignment with existing modalities. We believe that MIITA offers broad applicability to multimodal learning tasks involving tabular data, and its design principles can be extended to other domains like audio and video. MIITA represents a significant step forward in integrating dynamic tabular data with other modalities and opens new opportunities for future research in flexible, interpretable multimodal learning.

Acknowledgements

This work is supported by NSFC (grant nos. 62472102 to B.Y. and 62372117 to W.T.) and by the Natural Science Foundation of Shanghai (24ZR1490400 to W.T.). The computations in this research were performed using the CFFF platform of Fudan University.

References

[Arik and Pfister, 2020] Sercan O. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning, 2020.

[Bahri *et al.*, 2021] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. *arXiv preprint arXiv:2106.15147*, 2021.

[Bahri *et al.*, 2022] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption, 2022.

[Carballo *et al.*, 2023] Kimberly Villalobos Carballo, Liangyuan Na, Yu Ma, Léonard Boussioux, Cynthia Zeng, Luis R. Soenksen, and Dimitris Bertsimas. Tabtext: A flexible and contextual approach to tabular data representation, 2023.

[Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM, August 2016.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[Du *et al.*, 2024] Siyi Du, Shaoming Zheng, Yinsong Wang, Wenjia Bai, Declan P. O'Regan, and Chen Qin. Tip: Tabular-image pre-training for multimodal classification with incomplete data, 2024.

[Gorishniy *et al.*, 2021] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *CoRR*, abs/2106.11959, 2021.

[Gorishniy *et al.*, 2023] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data, 2023.

[Guo *et al.*, 2024] Y. Guo, S.D. Chen, J. You, et al. Multiplex cerebrospinal fluid proteomics identifies biomarkers for diagnosis and prediction of alzheimer's disease. *Nature Human Behaviour*, 8:2047–2066, 2024.

[Hager *et al.*, 2023] Paul Hager, Martin J. Menten, and Daniel Rueckert. Best of both worlds: Multimodal contrastive learning with tabular and imaging data, 2023.

[Hegselmann *et al.*, 2023a] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models, 2023.

[Hegselmann *et al.*, 2023b] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR, 2023.

[Higgins *et al.*, 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[Hollmann *et al.*, 2025] N. Hollmann, S. Müller, L. Purucker, et al. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025.

[Huang *et al.*, 2023] Jingmin Huang, Bowei Chen, Lan Luo, Shigang Yue, and Iadh Ounis. Dvm-car: A large-scale automotive dataset for visual marketing research and applications, 2023.

[Huang, 2023] Weichen Huang. Multimodal contrastive learning and tabular attention for automated alzheimer's disease prediction, 2023.

[Jack *et al.*, 2024] C.R. Jack, S.J. Andrews, T.G. Beach, et al. Revised criteria for the diagnosis and staging of alzheimer's disease. *Nature Medicine*, 30:2121–2124, 2024.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[Jr *et al.*, 2008] Clifford R. Jack Jr, Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[Kim and Mnih, 2019] Hyunjik Kim and Andriy Mnih. Disentangling by factorising, 2019.

[Kumar *et al.*, 2018] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations, 2018.

[Lu *et al.*, 2023] Jiaying Lu, Yongchen Qian, Shifan Zhao, Yuanzhe Xi, and Carl Yang. Mug: A multimodal classification benchmark on game data with tabular, textual, and visual fields. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 5332–5346. Association for Computational Linguistics, 2023.

[Mao *et al.*, 2023] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications, 2023.

[Nam *et al.*, 2023] Jaehyun Nam, Jihoon Tack, Kyungmin Lee, Hankook Lee, and Jinwoo Shin. Stunt: Few-shot tabular learning with self-generated tasks from unlabeled tables, 2023.

[O'Shea and Nash, 2015] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015.

[Potdar *et al.*, 2017] Kedar Potdar, Taher Pardawala, and Chinmay Pai. A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 175:7–9, 10 2017.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[Shi *et al.*, 2021] Xingjian Shi, Jonas Mueller, Nick Erickson, Mu Li, and Alex Smola. Multimodal autoML on structured tables with text fields. In *8th ICML Workshop on Automated Machine Learning (AutoML)*, 2021.

[Soenksen *et al.*, 2022] Luis R. Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M. Wiberg, Michael L. Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digital Medicine*, 5(1), September 2022.

[Somepalli *et al.*, 2021] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training, 2021.

[Su *et al.*, 2024] Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, Haoxuan Lan, Jiaming Tian, Jing Yuan, Junbo Zhao, Junlin Zhou, Kaizhe Shou, Liangyu Zha, Lin Long, Liyao Li, Pengzuo Wu, Qi Zhang, Qingyi Huang, Saisai Yang, Tao Zhang, Wentao Ye, Wufang Zhu, Xiaomeng Hu, Xijun Gu, Xinjie Sun, Xiang Li, Yuhang Yang, and Zhiqing Xiao. Tablegpt2: A large multimodal model with tabular data integration, 2024.

[Sudlow *et al.*, 2015] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, and et al. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779, 2015.

[van den Oord *et al.*, 2019] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[Wang and Sun, 2022] Zifeng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables, 2022.

[Wang *et al.*, 2024a] Laijun Wang, Zhiwei Cui, Shi Dong, and Ning Wang. Airport visibility classification based on multimodal fusion of image-tabular data. *IEEE Access*, 12:155082–155097, 2024.

[Wang *et al.*, 2024b] Zifeng Wang, Chufan Gao, Cao Xiao, and Jimeng Sun. Meditab: Scaling medical tabular data predictors via data consolidation, enrichment, and refinement, 2024.

[Wu *et al.*, 2021] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. Rethinking infonce: How many negative samples do you need?, 2021.

[Yi *et al.*, 2023] F. Yi, H. Yang, D. Chen, et al. Xgboost-shap-based interpretable diagnostic framework for alzheimer's disease. *BMC Medical Informatics and Decision Making*, 23(1):137, 2023.

[Zhang *et al.*, 2024] Han Zhang, Xumeng Wen, Shun Zheng, Wei Xu, and Jiang Bian. Towards foundation models for learning on tabular data, 2024.