M^3-VRD: Multimodal Multi-task Multi-teacher Visually-Rich Form Document Understanding

Anonymous ACL submission

Abstract

This paper presents a groundbreaking multimodal, multi-task, multi-teacher joint-grained knowledge distillation model for visually-rich form document understanding. The model is designed to leverage insights from both fine-grained and coarse-grained levels by facilitating a nuanced correlation between token and entity representations, addressing the complexities inherent in form documents. Additionally, we introduce new inter-grained and cross-grained loss functions to further refine diverse multi-teacher knowledge distillation transfer process, presenting distribution gaps and a harmonised understanding of form documents. Through a comprehensive evaluation across publicly available form document understanding datasets, our proposed model consistently outperforms existing baselines, showcasing its efficacy in handling the intricate structures and content of visually complex form documents.

1 Introduction

Understanding and extracting structural information from Visually-Rich Documents (VRDs), such as academic papers (Zhong et al., 2019; Ding et al., 2023b), receipts (Park et al., 2019), and forms (Jaume et al., 2019; Ding et al., 2023a), holds immense value for Natural Language Processing (NLP) tasks, particularly in information extraction and retrieval. While significant progress has been made in solving various VRD benchmark challenges, including layout analysis and table structure recognition, the task of form document understanding remains notably challenging. This complexity of the form document understanding arises from two main factors: 1) the involvement of two distinct authors in a form and 2) the integration of diverse visual cues. Firstly, forms mainly involve two primary authors: form designers and users. Form designers create a structured form to collect necessary information as a user interface. Unfortunately, the form layouts, designed to collect varied information, often lead to complex logical relationships, causing confusion for form users and heightening the challenges in form document understanding. Secondly, diverse authors in forms may encounter a combination of different document natures, such as digital, printed, or handwritten forms. Users may submit forms in various formats, introducing noise such as low resolution, uneven scanning, and unclear handwriting. Traditional document understanding models do not account for the diverse carriers of document versions and their associated noises, exacerbating challenges in understanding form structures and their components. Most VRD understanding models inherently hold implicit multimodal document structure analysis (Vision and Text understanding) knowledge either at fine-grained (Huang et al., 2022; Wang et al., 2022) or coarse-grained (Tan and Bansal, 2019; Li et al., 2019) levels. The fine-grained only models mainly focus on learning detailed logical layout arrangement, which cannot handle complex relationships of multimodal components, while the coarse-grained models tend to omit significant words or phrases. Hence, we introduce a novel joint-grained document understanding approach with multimodal multi-teacher knowledge distillation. This method leverages knowledge from various task-based teachers (pre-trained models) throughout the training process, intending to create more inclusive and representative multi- and joint-grained document representations.

Our contributions can be summarised as follows: 1) We present a groundbreaking multimodal, multi-task, multi-teacher joint-grained knowledge distillation model designed explicitly to understand visually-rich form documents. 2) The proposed model outperforms across publicly available form document datasets. 3) To the best of our knowledge, this research marks the first in adopting multi-task knowledge distillation, focusing on incorporating multimodal form document components.

Model	Modalities	Pre-training Datasets	Pre-training Tasks	Downstream Tasks	Granularity
Donut (2022)	V	IIT-CDIP	NTP	DC, VQA, KIE	Token
Pix2struct (2023b)	V	C4 corpus	NTP	VQA	Token
LiLT (2022)	T, S	IIT-CDIP	MVLM, KPL, CAI	DC, KIE	Token
BROS (2022)	T, S	IIT-CDIP	MLM, A-MLM	KIE	Token
LayoutLMv3 (2022)	T, S, V	IIT-CDIP	MLM, MIM, WPA	DC, VQA, KIE	Token
DocFormerv2 (2023)	T, S, V	IDL	TTL, TTG, MLM	DC, VQA, KIE	Token
Fast-StrucText (2023)	T, S, V	IIT-CDIP	MVLM, GTR, SOP, TIA	KIE	Token
FormNetV2 (2023a)	T, S, V	IIT-CDIP	MLM, GCL	KIE	Token
M^3-DVQA (Ours)	T, S, V	FUNSD, FormNLU	Multi-teacher Knowledge Distillation	KIE	Token, Entity

Table 1: Comparison with state-of-the-art models for receipt and form understanding. In the *Modalities* column, *T* represents Textual information, *V* represents Visual information, and *S* represents Spatial information.

2 Related Works

Visually Rich Document (VRD) understanding entails comprehending the structure and content of documents by capturing the underlying relations between textual and visual modalities. Several downstream tasks such as Key Information Extraction (KIE), Document Classification (DC), and Visual Question Answering (VQA), have contributed to raise the attention of the multimodal learning community. In this work, we cope with form documents, whose structure and content are particularly challenging to understand (Srivastava et al., 2020). Form documents possess intricate structures, involving collaboration between form designers, who craft clear structures for data collection, and form users, who interact with the forms based on their comprehension, with clarity and ease of understanding being variable. Table 1 enumerates the main characteristics of the latest models proposed for document understanding.

Vision-only approaches: They exclusively rely on the visual representation (denoted by *V* modality in Table 1) of the document components thus circumventing the limitations of state-of-the-art text recognition tools (e.g., Donut (Kim et al., 2022) and Pix2struct (Lee et al., 2023b)). Their document representations are commonly pre-trained using a Next Token Prediction (NTP) strategy, offering alternative solutions to traditional techniques based on Natural Language Processing.

Multimodal approaches: They leverage both the recognised text and the spatial relations (denoted by *T* and *S*) between document components (see, for instance, LiLT (Wang et al., 2022) and BROS (Hong et al., 2022)). The main goal is to complement raw content understanding with layout information. Expanding upon this multimodal framework, models such as LayoutLMv3 (Huang et al., 2022), DocFormerv2 (Appalaraju et al., 2023), Fast-StrucText (Zhai et al., 2023), and, FormNetV2 (Lee et al., 2023a) integrate the visual modality with text and layout information. These approaches are capable of capturing nuances in the document content hidden in prior works. To leverage multimodal relations, these models are typically pre-trained in a multi-task fashion, exploiting a curated set of token- or word-based pre-training tasks, such as masking or alignment.

Our approach aligns with the multimodal model paradigm, distinguishing itself by eschewing generic pre-training tasks reliant on masking, alignment, or NTP. Instead, it leverages the direct extraction of knowledge from *multiple teachers*, each trained on downstream datasets, encompassing *both entity and token levels* of analysis with the proposed inter-grained and cross-grained losses. This enriches the depth of understanding in visual documents, capturing intricate relationships and semantic structures beyond individual tokens.

3 Methodology

As previously noted, our paper focuses on interpreting visually rich documents, particularly form documents created and used collaboratively by multiple parties. To accomplish this objective, we introduce and employ two tiers of multimodal information: fine-grained and coarse-grained levels, which play a crucial role in understanding the structure and content of an input form page. Note that existing pre-trained visual-language models, whether designed for generic documents, possess implicit knowledge on either fine-grained or coarse-grained aspects. Hence, we propose an approach that harnesses knowledge from diverse pre-trained models throughout training. This strategy aims to generate

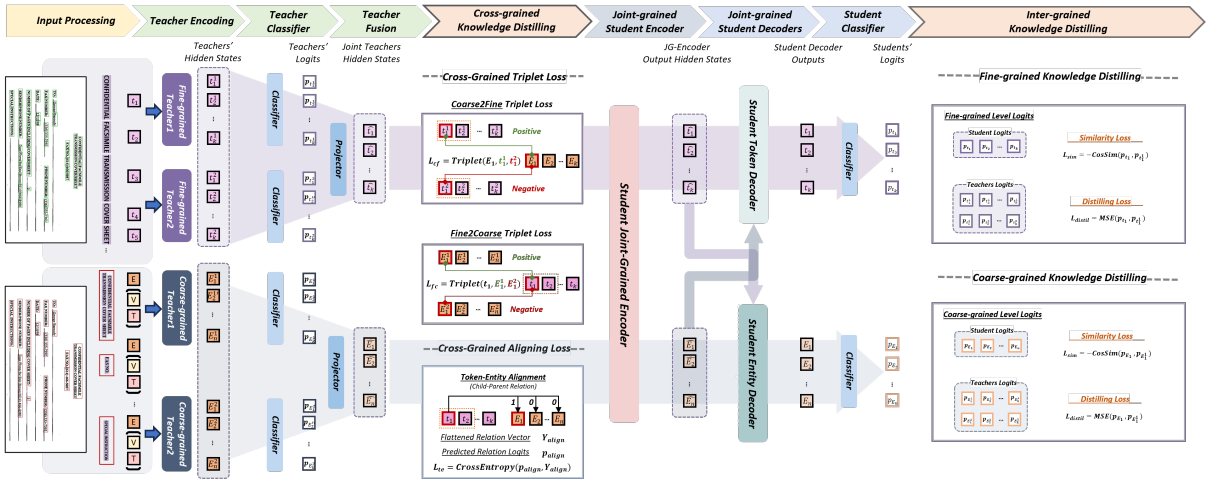


Figure 1: Multimodal Multi-task Multi-Teacher Joint-Grained Document Understanding Framework. Each section is aligned with the specific colours, Green: Section 3.2.1, Blue: Section 3.2.2, Orange: Section 3.3

more comprehensive and representative multi- and joint-grained document representations, ultimately enhancing the effectiveness of downstream tasks related to document understanding.

3.1 Preliminary Definitions

Prior to going through our proposed approach in detail, we would provide formal definitions for the terminology employed throughout this paper. We believe establishing clear and precise definitions could contribute to a comprehensive understanding of the concepts and terms integral to our research.

1) Fine-grained Document Understanding (Huang et al., 2022; Wang et al., 2022; Hong et al., 2022) is a pivotal aspect of document analysis, involving frameworks that offer detailed insights to comprehend document content, particularly when addressing token-level tasks, such as span-based information extraction and question answering. Regarding ***input features***, existing pre-trained models at the fine-grained level harness multimodal features, such as positional information and image-patch embedding, to enhance the fine-grained token representations. The ***pre-training phase*** incorporates several learning techniques, including Masked Visual-Language Modelling, Text-Image Matching, and Multi-label Document Classification, strategically designed to acquire inter or cross-modality correlations and contextual knowledge. However, it is essential to acknowledge the ***limitations*** of fine-grained frameworks, as their primary focus lies in learning the logical and layout arrangement of input documents. These frameworks may encounter challenges in handling complex multimodal components.

2) Coarse-grained Document Understanding (Tan and Bansal, 2019; Li et al., 2019) is a vital component in document analysis, with frameworks adept at grasping the logical relations and layout structures within input documents. Particularly well-suited for tasks like document component entity parsing, coarse-grained models excel in capturing high-level document understanding. Despite the dominant trend of fine-grained document understanding models, some research recognises (Tan and Bansal, 2019; Li et al., 2019) that the knowledge from general domain-based Visual-Language Pre-trained Models (VLPMs) could be leveraged to form a foundational document understanding. However, the coarse-grained document understanding models have significant ***limitations***, including their tendency to overlook detailed information, leading to the omission of significant words or phrases. Preliminary entity-level annotations are often necessary, and the current backbone models are pre-trained on the general domain, highlighting the need for document domain frameworks specifically pre-trained at the coarse-grained level.

3.2 Multimodal Multi-task Multi-teacher Joint-grained Document Understanding[1]

Therefore, we introduce a joint-grained document understanding framework $\mathcal{F}_{jg}$, designed to harness pre-trained knowledge from both fine-grained and coarse-grained levels. Our approach integrates in-

[1] Subsections are aligned with different colour in Figure 1, Green: Section 3.2.1, Blue: Section 3.2.2, Orange: Section 3.3

sights from multiple pre-trained backbones, facilitating a unified understanding of document content encompassing detailed nuances and high-level structures. It aims to synergise the strengths of fine-grained and coarse-grained models, enhancing the overall effectiveness of form understanding tasks.

3.2.1 Multimodal Multi-task Multi-Teacher

To facilitate this joint-grained framework, we employ **M**ultimodal **M**ulti-teachers from two **M**ulti-tasks, fine-grained and coarse-grained tasks within our framework. While the fine-grained teacher $\mathcal{F}_{fg}$ is characterised by checkpoints explicitly fine-tuned for the **token classification**, the coarse-grained teacher $\mathcal{F}_{cg}$ utilises fine-tuning checkpoints for the document component **entity classification**. The details of fine-grained and coarse-grained teacher models are articulated in Section 4.3. The ablation study of those teacher models is in Section 5.2. $\mathcal{F}_{fg}$ and $\mathcal{F}_{cg}$ get the encoded inputs of token and entity level, respectively, to acquire the corresponding last layer hidden states and logits for downstreaming procedures. For example, after feeding the sequence of tokens $\widetilde{\mathbf{t}} = \{\widetilde{t_1}, \widetilde{t_2}, ..., \widetilde{t_k}\}$ and sequence of multimodal entity embeddings $\widetilde{\mathbf{E}} = \{\widetilde{E_1}, \widetilde{E_2}, ..., \widetilde{E_n}\}$ into $\mathcal{F}_{fg_1}$ and $\mathcal{F}_{cg_1}$, respectively, we acquire the hidden states $\mathbf{t^1} = \{t_1^1, t_2^1, ..., t_k^1\}$ and $\mathbf{E^1} = \{E_1^1, E_2^1, ..., E_n^1\}$, as well as classification logits $\mathbf{p_{t^1}} = \{p_{t_1^1}, p_{t_2^1}, ..., p_{t_k^1}\}$ and $\mathbf{p_{E^1}} = \{p_{E_1^1}, p_{E_2^1}, ..., p_{E_n^1}\}$. Supposing $\mathbb{T} = \{\mathbf{t^1}, \mathbf{t^2}, ...\}$ and $\mathbb{E} = \{\mathbf{E^1}, \mathbf{E^2}, ...\}$ are hidden states from multiple teachers, the combined representations are fed into corresponding projection layers $\mathcal{L}_{fg}$ and $\mathcal{L}_{cg}$ to get the multi-teacher representations $\widehat{\mathbf{t}} = \{\widehat{t_1}, \widehat{t_2}, ..., \widehat{t_k}\}$ and $\widehat{\mathbf{E}} = \{\widehat{E_1}, \widehat{E_2}, ..., \widehat{E_n}\}$ for each grain.

3.2.2 Joint-Grained Learning

Our joint-grained learning framework comprises Joint-grained Encoder and Decoders.

The joint-grained encoder $\mathcal{E}$, implemented as a transformer encoder, is designed to learn the contextual correlation between fine-grained $\widehat{\mathbf{t}}$ and coarse-grained $\widehat{\mathbf{E}}$ representations. This enables the model to capture nuanced details at the token level while simultaneously grasping the high-level structures represented by entities within the document.

The joint-grained decoders $\mathcal{D}$ play a crucial role in processing the augmented joint-grained representations. For the fine-grained decoder $\mathcal{D}_{fg}$, the input comprises fine-grained token representations $\widehat{\mathbf{t}}$, with the entity representation serving as memory $\widehat{\mathbf{E}}$. This configuration allows the decoder to focus on refining and generating augmented token representations $\mathbf{t}$ based on the contextual information provided by both token and entity representations. In contrast, for coarse-grained decoder $\mathcal{D}_{cg}$, the input is the entity representation $\widehat{\mathbf{E}}$, while the memory consists of token representations $\widehat{\mathbf{t}}$. This approach enables the coarse-grained decoders to emphasise broader structures and relationships at the entity level, leveraging the memory of fine-grained token information to generate a more comprehensive entity representation $\mathbf{E}$. Overall, the proposed joint-grained architecture facilitates a comprehensive understanding of document content by incorporating fine-grained and coarse-grained perspectives.

The pre-training of different teacher models involves diverse techniques and features, so a simplistic approach of merely concatenating or pooling hidden states may not fully leverage the individual strengths of each model. Traditional self-/cross attention-based transformer encoders or decoders might encounter challenges in integrating knowledge from various grains, potentially introducing noise to specific grained weights. To address this concern, we propose using multiple types of losses to thoroughly explore implicit knowledge within the diverse teachers (pre-trained models).

3.3 Multimodal Multi-task Multi-Teacher Knowledge Distillation

This section introduces the multi-loss strategy to enhance inter-grained and cross-grained knowledge exchange, ensuring a more nuanced and effective integration of insights from fine-grained and coarse-grained representations. The accompanying multi-loss ablation study (Section 5.3) aims to optimise the synergies between multiple teacher models, thereby contributing to a more robust and comprehensive joint-grained learning process.

3.3.1 Task-oriented Cross Entropy Loss

The Task-oriented Cross Entropy (CE) loss is pivotal in facilitating a task-based knowledge distillation strategy. This is computed by comparing the predictions of the student model with the ground truth for each specific task. Adopting the CE loss provides the student model with direct supervisory signals, thereby aiding and guiding its learning process. Note that we address two task-oriented CE losses within our proposed approach, one from the token classification task and the other from the entity classification task. The output hidden

states from $\mathcal{D}_{fg}$ and $\mathcal{D}_{cg}$ are fed into classifiers to get the output logits $\mathbf{p_t} = \{p_{t_1}, p_{t_2}, ..., p_{t_k}\}$ and $\mathbf{p_E} = \{p_{E_1}, p_{E_2}, ..., p_{E_n}\}$. Supposing the label sets for fine-grained and entity-level tasks are $\mathbf{Y_t} = \{y_{t_1}, y_{t_2}, ..., y_{t_k}\}$ and $\mathbf{Y_E} = \{y_{E_1}, y_{E_2}, ..., y_{E_n}\}$, the fine-grained and coarse-grained Task-oriented Cross Entropy losses l_t and l_E are calculated as:

$$l_t = CrossEntropy(\mathbf{p_t}, \mathbf{Y_t}) \quad (1)$$

$$l_e = CrossEntropy(\mathbf{p_E}, \mathbf{Y_E}) \quad (2)$$

3.3.2 Inter-Grained Loss Functions

Since various pre-trained models provide different specific knowledge to understand the form comprehensively, effectively distilling valuable information from selected fine-tuned checkpoints may generate more representative token representations. In addressing this, we introduce two target-oriented loss functions tailored to distil knowledge from teachers at different levels. These aim to project the label-based distribution from fine-grained $\mathbf{p}_{\mathbb{T}} = \{\mathbf{p_{t^1}}, \mathbf{p_{t^2}}, ...\}$ or coarse-grained teacher logits $\mathbf{p}_{\mathbb{E}} = \{\mathbf{p_{E^1}}, \mathbf{p_{E^2}}, ...\}$ to corresponding student logits $\mathbf{p_t}$ and $\mathbf{p_E}$, enabling efficient learning of label distributions.

Similarity Loss: This is introduced as an effective method to distil knowledge from the output logits $\mathbf{p_t}$ and $\mathbf{p_E}$ of selected fine-grained or coarse-grained teacher checkpoints from $\mathbf{p}_{\mathbb{T}}$ and $\mathbf{p}_{\mathbb{E}}$. It aims to mitigate the logit differences between the student classifier and the chosen teachers using cosine similarity ($CosSim$), promoting a more aligned understanding of the label-based distribution. Supposing we have n_t and n_e teachers for fine-grained and coarse-grained tasks, respectively, the similarity loss of fine-grained l_{sim_t} and coarse-grained l_{sim_e} can calculated by:

$$l_{sim_t} = -\sum_{i}^{i=n_t}\sum_{j}^{j=k} CosSim(p_{t_j^i}, p_{t_j}) \quad (3)$$

$$l_{sim_e} = -\sum_{i}^{i=n_e}\sum_{j}^{j=n} CosSim(p_{E_j^i}, p_{E_j}) \quad (4)$$

Distilling Loss: Inspired by (Phuong and Lampert, 2019), we adopt an extreme logit learning model for the distilling loss. This loss implements knowledge distillation using Mean Squared Error (MSE) between the students' logits $\mathbf{p_t}$ and $\mathbf{p_E}$ and the teachers' logit sets $\mathbf{p}_{\mathbb{T}}$ and $\mathbf{p}_{\mathbb{E}}$. This method is employed to refine the knowledge transfer process further, promoting a more accurate alignment between the student and teacher models.

$$l_{distil_t} = \frac{1}{k}\sum_{j}^{j=k} MSE(p_{t_j^i}, p_{t_j}) \quad (5)$$

$$l_{distil_e} = \frac{1}{n}\sum_{j}^{j=n} MSE(p_{E_j^i}, p_{E_j}) \quad (6)$$

The introduction of these inter-grained loss functions, including the similarity loss and the distilling loss, contributes to mitigating distribution gaps and fostering a synchronised understanding of the form across various levels of granularity.

3.3.3 Cross-Grained Loss Functions

In addition, we incorporate cross-grained loss functions. While fine-grained and coarse-grained information inherently align, the joint-grained framework employs self-attention and cross-attention to approximate the correlation between token and entity representations. $\mathbb{T}$ and $\mathbb{E}$ are teachers hidden states sets, each $\mathbf{t^i} \in \mathbb{T}$ and $\mathbf{E^i} \in \mathbb{E}$ are represented $\mathbf{t^i} = \{t_1^i, t_2^i, ..., t_k^i\}$ and $\mathbf{E^i} = \{E_1^i, E_2^i, ..., E_n^i\}$ and $\mathbf{t}$ and $\mathbf{E}$ are hidden states from student decoder.

Cross-grained Triplet Loss: Inherent in each grained feature are parent-child relations between tokens and aligned semantic form entities. The introduction of triplet loss aids the framework in automatically selecting more representative feature representations by measuring the feature distance from one grain to another-grained aligned representation. This effectively enhances joint-grained knowledge transfer, optimising the overall understanding of the form. For acquiring the loss $l_{triplet_{fg}}$ to select fine-grained teachers based on coarse-grained distribution adaptively, we define the anchor as each entity $E_i \in \mathbf{E}$ which has the paired token representations $t_i^1 \in \mathbf{t^1}$ and $t_i^2 \in \mathbf{t^2}$ (if the number of teachers is more significant than 2, randomly select two of them). The L-2 norm distance is used to measure the distance between fine-grained teachers (t_i^1, t_i^2) and anchor E_j, where the more similar entities are treated as positive samples (t_i^{pos}) otherwise negative (t_i^{neg}). For coarse-grained triplet loss $l_{triplet_{cg}}$, the same measurements are adopted for coarse-grained teacher positive (E_j^{pos}) and negative selection (E_j^{neg}) for an anchor t_i. Supposing the j-th, $l_{triplet_{fg}}$ and $l_{triplet_{cg}}$ are defined:

$$l_{triplet_{fg}} = \frac{1}{k}\sum_{i}^{i=k} Triplets(E_j, t_i^{pos}, t_i^{neg}) \quad (7)$$

$$l_{triplet_{cg}} = \frac{1}{k}\sum_{i}^{i=k} Triplets(t_i, E_j^{pos}, E_j^{neg}) \quad (8)$$

As one entity is typically paired with more than one token, when calculating $l_{triplet_{cg}}$, we will consider all k entity-token pairs.

Cross-grained Alignment Loss: In addition to the triplet loss, designed to filter out less representative teachers, we introduce another auxiliary task. This task focuses on predicting the relations between tokens and entities, providing an additional layer of refinement to the joint-grained framework. The cross-grained alignment loss further contributes to the comprehensive learning and alignment of token and entity representations, reinforcing the joint-grained understanding of the form document. For an input form document page containing k tokens and n entities, we have a targeting tensor $\mathbf{Y_{align}}$ where $Dim(\mathbf{Y_{align}}) = \mathbb{R}^{k \times n}$. We use acquired alignment logit $\mathbf{p_{align}} = \mathbf{t} \times \mathbf{E}$ to represent the predicted token-entity alignments. The cross-grained alignment loss l_{align} can be calculated by:

$$l_{align} = CrossEntropy(\mathbf{p_{align}}, \mathbf{Y_{align}}) \quad (9)$$

4 Evaluation Setup

4.1 Datasets[2]

FUNSD (Jaume et al., 2019) comprises 199 noisy scanned documents from various domains including marketing, advertising, and science reports, related to US tobacco firms. It is split into train and test sets (149/50 documents), and each document is presented in either printed or handwritten format with low resolutions. Our evaluation focuses on the semantic-entity labeling task that identifies four predefined labels (i.e., question, answer, header, and other) based on input text content.

FormNLU (Ding et al., 2023a) consists of 867 financial form documents collected from Australian Stock Exchange (ASX) filings. It includes three form types: digital ***(D)***, printed ***(P)***, and handwritten ***(H)***, and is split into five sets: train-***D*** (535), val-*D* (76), test-*D* (146), test-***P*** (50), and test-***H*** (50 documents) and supports two tasks: Layout Analysis and Key Information Extraction. Our evaluation focuses on the layout analysis that identifies seven labels (i.e., title, section, form key, form value, table key, table value, and others), detecting each document entity, especially for ***P*** and ***H***, the complex multimodal form document.

4.2 Baselines and Metrics

For **token-level information extraction** baselines, we use three Document Understanding (DU) models: LayoutLMV3 (Huang et al., 2022), LiLT (Wang et al., 2022), and BROS (Hong et al., 2022). LayoutLMV3 employs a word-image patch alignment, that utilises a document image along with its corresponding text and layout position information. In contrast, LiLT and BROS focus only on text and layout information without incorporating images. LiLT uses a bi-directional attention mechanism across token embedding and layout embedding, whereas BROS uses a relative spatial encoding between text blocks. For **entity-level information extraction** baselines, we use two vision-language (VL) models: LXMERT (Tan and Bansal, 2019) and VisualBERT (Li et al., 2019). Compared to the two DU models, these VL models use both image and text input without layout information. LXMERT focuses on cross-modality learning between word-level sentence embeddings and object-level image embeddings, while VisualBERT simply inputs image regions and text, relying on implicit alignments within the network. For **evaluation metrics**, inspired by (Jaume et al., 2019) and (Ding et al., 2023a), we primarily use F1-score to represent both overall and detailed performance breakdowns, aligning with other baselines.

4.3 Implementation Details[3]

In token-level experiments, we fine-tuned LayoutLMV3-base using its text tokeniser and image feature extractor. We also fine-tuned LiLT combined with RoBERTa base. In entity-level experiments, we employ pre-trained BERT (748-d) for encoding textual content, while ResNet101(2048-d) is used for region-of-interest(RoI) feature to capture the visual aspect. These extracted features serve as input for fine-tuning LXMERT and VisualBERT. All fine-tuned models serve as teacher models. Our hyperparameter testing involves a maximum of 50 epochs with learning rates set at 1e-5 and 2e-5. All are conducted on a Tesla V100-SXM2 with 16GB graphic memory and 51 GB memory, CUDA 11.2.

5 Results

5.1 Overall Performance

Extensive experiments are conducted to highlight the effectiveness of the proposed **Multimodal Multi-task Multi-Teacher** framework, including *joint-grained learning*, *multi-teacher* and *multi-loss* architecture. Table 2 shows representative

[2]The statistics of token/entity are shown in Table 5 and 6.

[3]Additional Implemtnation Details are in Appendix D

Model	Config & Loss	FUNSD	FormNLU	
			P	***H***
BROS	Single Teacher	82.44	92.45	93.68
LiLT	Single Teacher	87.54	96.50	91.35
LayoutLMv3	Single Teacher	90.61	95.99	97.39
JG-$\mathcal{E}$	Joint Cross Entropy	90.45	94.91	96.55
JG-$\mathcal{D}$	Joint Cross Entropy	90.48	95.68	97.62
JG-$\mathcal{E}$&$\mathcal{D}$	Joint Cross Entropy	90.57	95.93	97.62
MT-JG-$\mathcal{E}$&$\mathcal{D}$ (Ours)	Joint Cross Entropy	90.53	97.21	97.75
	+ *Sim*	**91.05**	98.25	98.09
	+ *Distil*	90.90	98.12	97.72
	+ *Triplet*	90.28	97.58	97.28
	+ *Align*	90.55	97.24	97.42
	+ *Sim* + *Distil* + *Triplet* + *Align*	90.92	**98.69**	**98.39**

Table 2: Overall Performance of Model and Configurations on FormNLU printed ***P*** and handwritten ***H***. The full form of acronyms can be found in Section 5.1. The best is in **bold**. The second best is underlined.

FG Teacher	CG Teacher	FUNSD	FormNLU	
			P	**H**
LLmv3	VBERT	90.19	94.72	96.99
	LXMERT	**90.57**	95.93	97.62
	Transformer	90.22	93.65	95.94
LiLT	VBERT	87.66	**97.65**	90.53
	LXMERT	87.34	96.76	91.18
	Transformer	87.91	97.20	90.58
LLmv3	VBERT&LXMERT	90.42	95.05	97.25
LLmv3 & LiLT	LXMERT	90.39	96.73	97.42
LLmv3&LiLT	**VBERT&LXMERT**	90.53	97.21	**97.75**

Table 3: Comparison of Performance across Teacher Combinations. FG: Fine-Grained, CG: Coarse-Grained, LLmv3: LayoutLMv3, VBERT: VisualBERT. The best is in **bold**. The second best is underlined. This ablation study is based on only Joint Cross Entropy Loss.

model configurations on various adopted modules. LayoutLMv3 performs notably superior to BROS and LiLT, except for the FormNLU printed test set. LayoutLMv3 outperforms around 3% and 4% the second-best baseline on FUNSD and FormNLU handwritten sets, respectively. This superiority can be attributed to LayoutLMv3's utilisation of patched visual cues and textual and layout features, resulting in more comprehensive multimodal representations. So we found LayoutLMv3 would be a robust choice for fine-grained baselines in further testing[4]. To find the most suitable **Joint-Grained learning** (JG), we compare the results of single-teacher joint-grained frameworks including Encoder ($\mathcal{E}$) only, Decoder ($\mathcal{D}$) only, and Encoder with Decoder ($\mathcal{E}$&$\mathcal{D}$). Table 2 illustrates $\mathcal{E}$&$\mathcal{D}$ achieving the highest performance among three baselines. However, upon integrating multiple teachers from each grain (MT-JG-$\mathcal{E}$&$\mathcal{D}$), competitive performance is observed compared to the baselines on both FormNLU printed (***P***) (from LiLT 96.5% to 97.21%) and handwritten set (***H***) (from LiLT 97.39% to 97.75%). Still, additional techniques may be necessary to distil the cross-grained multi-teacher information better.

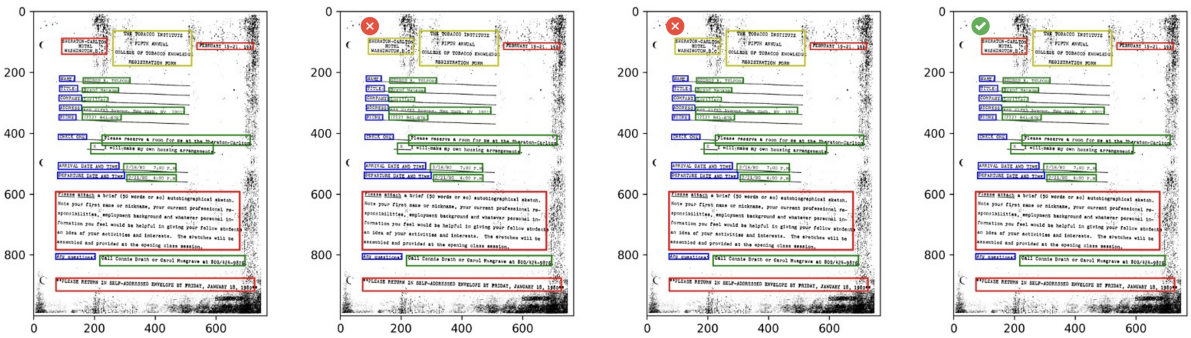
To thoroughly distil joint-grained knowledge from multiple teachers, we introduced multiple loss functions encompassing **Multiple auxiliary tasks**. These functions capture teacher knowledge from inter-grained and cross-grained perspectives, generating representative token embeddings. Typically, using either inter-grained or coarse-grained loss individually leads to better performance than the best baselines across various test sets. Inter-grained Similarity (*Sim*) and Distilling (*Distil*) loss consistently achieve higher F1 scores in nearly all test sets. Moreover, cross-grained *Triplet* and alignment (*Align*) losses outperform the best baseline on the FormNLU (***P***) or (***H***). This highlights the effectiveness of the proposed multi-task learning approach in enhancing token representations by integrating knowledge from joint-grained multi-teachers. Inter-grained loss functions exhibit higher robustness on both datasets, whereas cross-grained loss functions only perform well on FormNLU. This difference may stem from the FUNSD being sourced from multiple origins, whereas FormNLU is a single-source dataset. Coarse-grained loss functions may excel on single-source documents by capturing more prevalent knowledge but might introduce noise when applied to multiple sources. Also, the model demonstrates its most competitive performance by integrating all proposed loss functions (+*Sim*+*Distil*+*Triplet*+*Align*). This highlights how the proposed inter-grained and cross-grained loss functions enhance multi-teacher knowledge distillation in form understanding tasks[5].

5.2 Effect of Multi-Teachers

We analysed various teacher combinations to ensure they provide sufficient knowledge for improving joint-grained representations, as depicted in Table 3. For fine-grained teachers, since BROS underperforms compared to others, we only include the performance of its counterparts. The LayoutLMv3-based joint framework performs better, outperforming LiLT-based by approximately 3% on FUNSD and over 5% on FormNLU (***H***).

[4]We chose LLmv3 and LXMERT for JG and select LLMv3&LilT and VBERT&LXMERT for MT-JG-$\mathcal{E}$&$\mathcal{D}$. More teacher combinations analysis is in Section 5.2.

[5]More loss combination analysis is in Section 5.3



(a) **Ground Truth** (b) **JG-$\mathcal{E}\&\mathcal{D}$** (c) **LayoutLMv3** (d) **Ours**

Figure 2: Example output showing (a) Ground Truth (b) JG-$\mathcal{E}\&\mathcal{D}$ (c) LayoutLMv3, and (d) Ours on a FUNSD page. The color code for layout component labels is as follows; Question, Answer, Header, Other. Our model, employing the best loss combination (cross-entropy + similarity) on FUNSD, accurately classified all layout components.

Loss Functions				FUNSD	FormNLU	
Similarity	Distiling	Triplet	Alignment		*P*	*H*
O	X	X	X	**91.05**	98.25	98.09
X	O	X	X	90.90	98.12	97.72
X	X	O	X	90.28	97.58	97.28
X	X	X	O	90.55	97.24	97.42
O	O	X	X	90.63	98.53	97.22
O	X	O	X	90.51	97.71	97.79
O	X	X	O	90.82	97.80	98.05
X	O	O	X	90.82	98.22	98.35
X	O	X	O	90.83	98.63	97.45
O	O	O	X	90.79	98.56	97.72
O	O	X	O	90.66	**98.72**	97.85
O	O	O	O	90.92	98.69	**98.39**

Table 4: Performance comparison across loss functions. The best is in **bold**. The second best is underlined.

This improvement can be attributed to the contextual learning facilitated by visual cues. Notably, LiLT achieves the highest performance on the FormNLU (***P***), likely due to its well-designed positional-aware pre-training tasks. For coarse-grained teachers, pre-trained backbones demonstrate better robustness than randomly initialised Transformers, highlighting the benefits of general domain pre-trained knowledge in form understanding tasks. Table 3 illustrates multiple teachers cannot always ensure the best performance, however, the robustness of the proposed model is enhanced by capturing more implicit knowledge from cross-grained teachers.

5.3 Effect of Loss Functions

To comprehensively investigate the impact of different loss functions and their combinations, we present the performance of various combinations in Table 4. While employing inter-grained loss individually often proves more effective than using cross-grained loss alone, combining the two losses can enhance knowledge distillation from joint-grained multi-teachers. For instance, concurrently employing distilling(Distil) and Triplet loss improved accuracy from 97.72% to 98.35%. Notably, stacking all proposed loss functions resulted in the best or second-best performance across all test sets, showcasing their effectiveness in distilling knowledge from multi-teacher to student models for generating more representative representations. Even though cross-grained Triplet and Alignment losses were ineffective individually, when combined with inter-grained loss, they significantly improved knowledge distillation effectiveness.

5.4 Qualitative Analysis: Case Studies[6]

We visualised the sample results for the top 3 - Our best model with the best configuration, the best baseline *LayoutLMv3* and the second best baseline *JG-$\mathcal{E}\&\mathcal{D}$* of FUNSD in Figure 2. We can see that both *LayoutLMv3* and *JG-$\mathcal{E}\&\mathcal{D}$* have wrongly recognised an *Other* (marked by a white cross in red circle), whereas ours has accurately recognised all document tokens and components.

6 Conclusion

We introduced a Multimodal Multi-task Multi-Teacher framework in Visually-Rich form documents. Our model incorporates *multi-teacher*, *multi-task*, and *multi-loss*, and the results show the robustness in capturing implicit knowledge from multi-teachers for understanding diverse form document natures, such as scanned, printed, and handwritten. We hope our work provides valuable insights into leveraging multi-teacher and multi-loss strategies for document understanding research.

[6]A Case Study for FormNLU can be found in Figure 3

Limitations

Benchmark Scope: Despite the paramount importance of document understanding across various domains such as finance, medicine, and resources, our study is constrained by the limited availability of visually-rich form document understanding datasets, particularly those of high quality. In this research, we solely rely on publicly available English-based form document understanding datasets. The scope of benchmark datasets, therefore, may not comprehensively represent the diversity and complexity present in form documents across different languages and industries.

Availability of Document Understanding Teachers: The current limitation stems from the reliance on general document understanding teacher models due to the absence of large pre-trained form-specific document models. The availability of high-quality teachers specifically tailored for form document understanding is crucial. Future advancements in the field would benefit from the development of dedicated pre-trained models for form document understanding, providing more accurate knowledge transfer during training.

References

Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. 2023. Docformerv2: Local features for document understanding. *arXiv preprint arXiv:2306.01733*.

Yihao Ding, Siqu Long, Jiabin Huang, Kaixuan Ren, Xingxiang Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023a. Form-nlu: Dataset for the form natural language understanding. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2807–2816.

Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023b. Pdf-vqa: A new dataset for real-world vqa on pdf documents. In *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track*, pages 585–601. Springer Nature Switzerland.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10767–10775.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 498–517. Springer.

Chen-Yu Lee, Chun-Liang Li, Hao Zhang, Timothy Dozat, Vincent Perot, Guolong Su, Xiang Zhang, Kihyuk Sohn, Nikolay Glushnev, Renshen Wang, Joshua Ainslie, Shangbang Long, Siyang Qin, Yasuhisa Fujii, Nan Hua, and Tomas Pfister. 2023a. FormNetV2: Multimodal graph contrastive learning for form document information extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9011–9026. Association for Computational Linguistics.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023b. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

Mary Phuong and Christoph Lampert. 2019. Towards understanding knowledge distillation. In *International conference on machine learning*, pages 5142–5151. PMLR.

Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey, and Snehasis Mukherjee. 2020. Visual question answering using deep learning: A survey and performance analysis. In *Computer Vision and Image Processing - 5th International Conference,CVIP 2020*, volume 1377 of *Communications in Computer and Information Science*, pages 75–86. Springer.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5100–5111.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7747–7757.

Mingliang Zhai, Yulin Li, Xiameng Qin, Chen Yi, Qunyi Xie, Chengquan Zhang, Kun Yao, Yuwei Wu, and Yunde Jia. 2023. Fast-structext: An efficient hourglass transformer with modality-guided dynamic token merge for document understanding. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023*, pages 5269–5277.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE.

A Statistics of tokens and entities

The following Table 5 and 6 demonstrates the number of tokens(length) and number of document entities. While FUNSD has 4 types(Question, Answer, Header, Other) of document entities, FormNLU has 7 types(Title, Section, Form Key, Form Value, Table Key, Table Value, Other). For the FormNLU, we applied two types of test set, including Printed **P** and Handwritten **H**.

FUNSD (Testing)	**Question**	**Answer**	**Header**	**Other**	**Total**
Entity	1077	821	122	312	**2332**
Token	2654	3294	374	2385	**8707**

Table 5: FUNSD Testing Dataset Distribution by Label.

FormNLU (Testing)		**Title**	**Section**	**Form Key**	**Form Value**	**Table Key**	**Table Value**	**Others**	**Total**
P	**Entity**	98	100	346	332	250	249	152	**1527**
H		100	100	348	315	249	226	149	**1487**
P	**Token**	700	1258	1934	1557	993	389	3321	**10152**
H		742	1031	1805	866	779	366	2918	**8507**

Table 6: FormNLU Testing Dataset Distribution by Label, where ***P*** and ***H*** are printed and handwritten sets.

B Breakdown Result Analysis

Model	**Config**	**Overall**	**Breakdown**		
			Header	**Question**	**Answer**
LiLT	Teacher	87.54	55.61	90.20	88.34
LayoutLMv3	Teacher	90.61	66.09	91.60	**92.78**
JG-$\mathcal{E}$	Joint CE	90.45	64.94	91.70	92.67
JG-$\mathcal{D}$	Joint CE	90.48	64.07	91.58	92.73
JG-$\mathcal{E}$&$\mathcal{D}$	Joint CE	90.57	64.66	91.48	92.73
MT-JG-$\mathcal{E}$&$\mathcal{D}$	Joint CE	90.53	61.24	92.40	91.75
	Sim	**91.05**	64.81	92.58	92.46
	Distil	90.90	**66.96**	**92.61**	91.97
	Triplet	90.28	62.44	92.00	91.44
	Align	90.55	63.81	91.82	92.29
	+Sim+Distil +Triplet+Align	90.92	64.22	92.54	92.31

Table 7: Breakdown Results of FUNSD dataset.

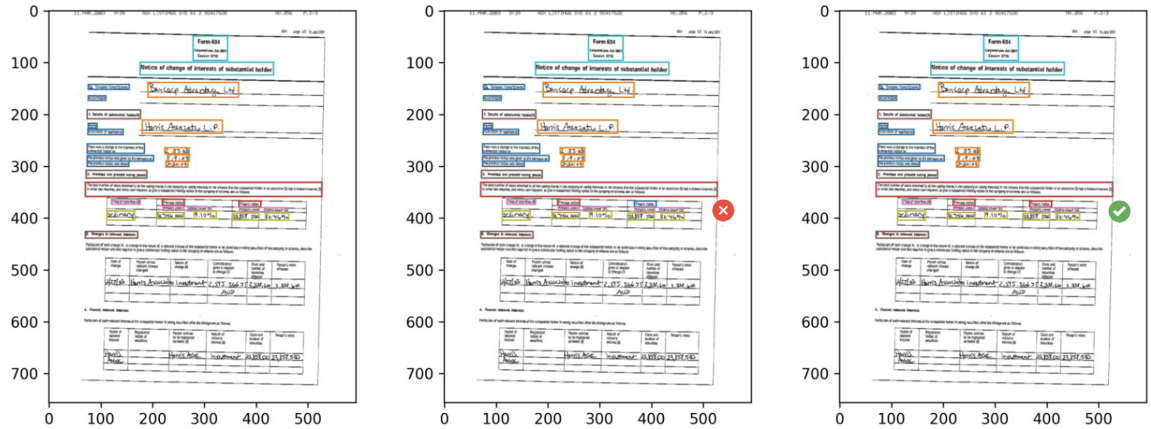
As shown in Table 7, for the FUNSD dataset, we could find all Joint-Grained(JG-) frameworks can have a delicate performance on recognising *Question* and *Answer*, but decreased in Header classification. This might result from the limited number of *Headers* in the FUNSD, leading to inadequate learning of the fine-grained and coarse-grained *Header* information. Multi-task-oriented inter-grained and coarse-grained functions can increase the performance of *Question* recognition by boosting the knowledge distilling from joint-grained multi-teachers. Especially, inter-grained knowledge distillation methods can achieve around 1% higher than LayoutLMv3. The FUNSD dataset cannot illustrate the benefits of cross-grained loss functions well.

For FormNLU printed and handwritten sets, the joint-grained framework and proposed loss functions can effectively improve *Section* (*Sec*) and Title recognition. As the *Title*, *Section* and *Form_key* (*F_K*) are normally located at similar positions for single-source forms, this may demonstrate both joint-grained framework and multi-task loss function could distil knowledge. Additionally, baseline models are not good at recognising table keys and values, especially handwritten sets. As we use the layoutLMv3 in the joint-grained framework, the performance of recognising table-related tokens is not good for the joint-learning framework. After integrating multiple teachers, the performance has increased from 91.97% to 97.35% on the printed set. The proposed multi-task loss functions may achieve a higher performance of 97.96%. Significant improvements can also be observed across two test sets across all table-related targets. This illustrates that the joint-grained multi-teacher framework can effectively tackle the limitation of one teacher to generate more comprehensive token representations, and the inter-grained and cross-grained loss could boost the effective knowledge exchange to make the generalisation and robustness of the entire framework.

C Additional Qualitative Analysis

In our qualitative evaluation, we took a closer look at the results by visualising the output of the top two models—our best-performing model with the optimal configuration and the baseline *LayoutLM3*—on the FormNLU handwritten set, as presented in Figure 3. This examination revealed a notable discrepancy between the models. Specifically, *LayoutLM3* exhibited an erroneous identification of the Table Key as a Form Key. In contrast, our model demonstrated a higher level of precision by accurately recognising and distinguishing all components within this intricate and noise-laden handwritten document.

This illustrative case serves as a compelling example highlighting the challenges associated with relying solely on knowledge from a single document to understand teachers. The complexity of distinguishing various document structures, such as the nuanced difference between a form key and a table key, becomes evident. The inadequacy of a singular teacher's knowledge in capturing such intri-



(a) **Ground Truth** (b) **LayoutLMv3** (c) **Ours**

Figure 3: Example output showing (a) Ground Truth (b) LayoutLMv3, and (c) Ours on a FormNLU handwritten test set. The color code for layout component labels is as follows; Title, Section, Form Key, Form Value, Table Key, Table Value, Other. Our model, the best loss combination (+Sim+Distil+Triplet+Align) on FormNLU **H**, accurately classified all layout components.

Model	Config	FormNLU Printed Overall and Breakdown							FormNLU Handwritten Overall and Breakdown						
		Overall	Sec	Title	F_K	F_V	T_K	T_V	Overall	Sec	Title	F_K	F_V	T_K	T_V
LiLT	Teacher	96.50	98.32	96.97	98.84	96.62	96.57	93.60	91.35	95.39	99.50	94.81	90.67	84.19	89.81
LayoutLMV3	Teacher	95.99	98.45	97.96	97.97	96.73	92.37	92.98	97.39	**99.33**	99.01	**99.85**	98.24	93.95	95.95
JG-$\mathcal{E}$	Joint CE	94.91	**99.66**	98.99	98.11	95.73	90.14	90.31	96.55	**99.33**	99.01	99.42	**98.56**	88.37	94.67
JG-$\mathcal{D}$	Joint CE	95.68	**99.66**	**100.00**	98.55	96.45	91.94	91.10	97.62	**99.33**	99.01	**99.85**	**98.56**	93.02	95.98
JG-$\mathcal{E}$&$\mathcal{D}$	Joint CE	95.93	**99.66**	97.96	97.82	97.18	91.97	92.15	97.62	**99.33**	99.01	**99.85**	98.40	93.74	95.75
MT-JG-$\mathcal{E}$&$\mathcal{D}$	Joint CE	97.21	99.32	98.48	99.57	96.58	97.35	95.06	97.75	97.67	99.50	99.13	97.93	95.55	96.41
	Sim	98.25	99.32	99.49	99.28	97.75	**97.96**	**97.12**	98.09	99.00	**100.00**	99.27	98.25	96.45	96.61
	Distil	98.12	99.32	**100.00**	**99.71**	97.90	97.55	96.30	97.72	97.35	**100.00**	99.13	97.62	95.75	97.07
	Triplet	97.58	99.32	99.49	99.28	97.18	97.55	95.87	97.28	98.00	**100.00**	98.83	97.31	93.90	96.83
	Align	97.24	99.32	98.48	**99.71**	96.57	96.13	95.47	97.42	**99.33**	99.50	99.13	96.85	92.86	97.52
	+Sim+Distil +Triplet+Align	**98.69**	99.32	**100.00**	**99.71**	**99.25**	97.35	**97.12**	**98.39**	98.33	**100.00**	99.56	98.09	**96.94**	**97.75**

Table 8: Overall and Breakdown Analysis of FormNLU Printed Set and Handwritten Set. The categories of FormNLU dataset Task A include Section (Sec), Title, Form_Key (F_K), Form_Value (F_V), Table_Key (T_K), Table_Value (T_V).

cacies emphasises the importance of our proposed **Multi-modal Multi-task Multi-Teacher** framework, which leverages insights from multiple teachers to enhance the robustness and accuracy of form document understanding.

D Additional Implementation Details

The table presented in Table 9 outlines the number of total parameters and trainable parameters across various model configurations. It is evident that the choice of teacher models primarily determines the total number of parameters. As the number of teachers increases, there is a corresponding enhancement in the total parameter count. Furthermore, the architecture of the student model significantly influences the number of trainable parameters. For instance, encoder-decoder-based student models exhibit a higher count of trainable parameters compared to architectures employing only an encoder or decoder. This discrepancy implies that training encoder-decoder models demands more computational resources. Despite the variation in trainable parameters among different student model architectures, it is noteworthy that the overall number remains substantially smaller than that of single-teacher fine-tuning processes. This observation underscores the efficiency of student model training in comparison to fine-tuning pre-trained models.

Fine-grained	Coarse-Grained	Configure	# Para	# Trainable
LiLT	N/A	Teacher	130,169,799	130,169,799
LayoutLMv3	N/A	Teacher	125,332,359	125,332,359
	LXMERT	JG-Encoder	393,227,514	19,586,415
		JG-Decoder	423,952,890	50,311,791
		JG-$\mathcal{E}$&$\mathcal{D}$	440,494,842	66,853,743
	VisualBERT&LXMERT		557,260,798	70,394,991
Layoutlmv3&LiLT	LXMERT		574,205,889	68,034,159
	VisualBERT&LXMERT		688,611,013	71,575,407

Table 9: Model configurations and parameters