Flow-Based Page Unique Semantic Mapping Architecture for Document Visual Question Answering

Haosen Wang[1], Jing Xiao[2], Chaochao Du[1], Xiaowang Zhang[1]*, Zhiyong Feng[1]
[1] School of Computer Software,Tianjin University,Tianjin,China
[2] International Engineering Institute,Tianjin University,Tianjin,China
{haosenwang,xiaojing_95,chuck,xiaowangzhang,zyfeng}@tju.edu.cn

Abstract

Document Visual Question Answering (DocVQA) aims to generate answers by jointly understanding the textual, layout, and visual elements within document images. Although end-to-end vision-based generative methods have reduced dependency on OCR, they still struggle to achieve precise evidence localization when page semantics are complex and highly similar. However, existing research lacks an in-depth theoretical analysis of the question-driven semantic representation space, failing to fundamentally address the distinguishability problem among semantically similar pages. To fill this theoretical gap, we propose and prove that, given a specific question, each page possesses a unique semantic representation, and there exists a bijective mapping between the page and its unique semantics. Based on this theoretical foundation, we introduce the **F**low-Based Page **U**nique Semantic **M**apping **A**rchitecture (**FUMA**), which reconstructs evidence localization from similarity-based retrieval into precise selection on unique semantics. FUMA employs fine-grained cross-modal attention to extract discriminative cues and utilizes flow-based reversible transformations with likelihood regularization to learn bijective mappings, ensuring that each page obtains a unique semantic representation. Moreover, a multi-expert collaboration mechanism complementarily models fine-grained multimodal information within each page, achieving robust answer generation. Experimental results demonstrate that FUMA significantly outperforms existing methods in both evidence localization and answer generation.

1 Introduction

Document Visual Question Answering (DocVQA) is a cross-modal artificial intelligence task (Kafle et al., 2018; Hudson and Manning, 2019) that aims to generate answers by understanding and reasoning over the text, layout structure, and visual elements in document images. As a key technique for document understanding and vision–language integration (Xu et al., 2020; Herzig et al., 2020; Qin et al., 2025; Pang et al., 2024), DocVQA plays an important role in intelligent office applications (Lee et al., 2022), information retrieval (Mathew et al., 2021; Wang et al., 2025), and human–computer interaction (Ding et al., 2022).

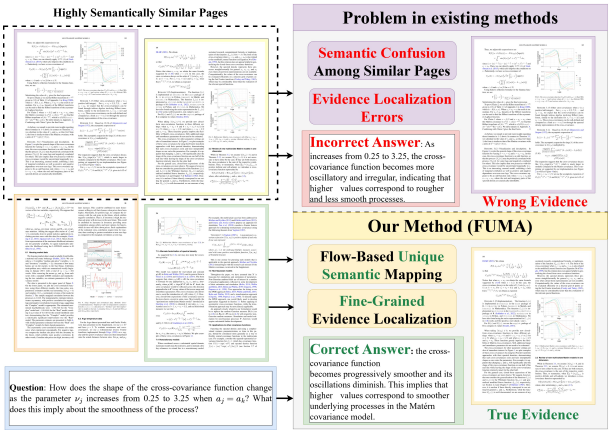

Figure 1: **Existing Methods vs. FUMA**. Existing methods rely on coarse-grained similarity matching, causing confusion among semantically similar pages (overlapping representations). FUMA establishes bijective mappings to unique page semantics via Flow-Based transformation, transforming evidence localization from **similarity-based retrieval** to **precise selection over a well-separated semantic space**.

For the DocVQA task, early studies primarily adopted OCR-based pipeline methods (Mathew et al., 2021; Hu et al., 2020; Xu et al., 2020), in which an OCR system is used to extract textual content and spatial layout information from documents, followed by text-understanding models for evidence localization and answer generation. However, this paradigm has clear limitations. First, model performance is strongly constrained

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36260–36280

by OCR quality, and recognition errors propagate through the pipeline and are amplified across stages (Kim et al., 2022). Second, in scenarios such as multi-page document retrieval, cross-page reasoning, and understanding complex visual elements, these methods often suffer from evidence-grounding errors, broken reasoning chains, and limited generalization (Cho et al., 2025). To overcome these limitations, recent research has shifted toward end-to-end vision-to-text generation (Lee et al., 2023), which weakens the explicit dependence on OCR by directly generating text from images and enables end-to-end modeling of visual structures. Although this direction has achieved substantial progress (Yu et al., 2025), it still faces key challenges in long-document and multi-page settings: because evidence-page localization relies on coarse-grained question–page semantic similarity matching, it is difficult to localize the correct page when pages are highly semantically similar; moreover, existing methods inadequately model fine-grained cross-modal features within a page, thereby limiting the accuracy and robustness of answer generation.

Therefore, in DocVQA, the challenge arises not only from the confusion among semantically similar pages, but also from a more fundamental issue: question-driven evidence localization requires a semantically separable representation space, where each page conditioned on the question exhibits both uniqueness and discriminability. If each page corresponds to a unique, question-specific semantics, different pages can be well separated in this semantic space; consequently, evidence-page localization reduces to a precise selection problem over unique semantics and provides stable, traceable grounding for subsequent answer generation. Accordingly, we propose and prove the following proposition: given a question, each document page admits a unique semantics with respect to that question, and there exists a bijection that maps document pages to their unique semantics. To satisfy semantic uniqueness and the invertibility implied by the bijection, we adopt an inherently invertible flow-based formulation to learn this mapping and propose the **F**low-Based Page **U**nique Semantic **M**apping **A**rchitecture (**FUMA**). On the one hand, we employ an attention mechanism to select fine-grained, question-relevant cross-modal cues within each page, allowing the model to retain evidence-discriminative details even among semantically similar pages. On the other hand, we introduce flow-based likelihood regularization to learn an invertible transformation path, alleviating representation collapse where multiple pages are mapped to similar semantics and ensuring that each page, under the question, is assigned a distinct and identifiable unique semantics. Furthermore, for answer generation grounded on the evidence page's unique semantics, we incorporate a collaborative mixture-of-experts mechanism to complementarily model and decode fine-grained multimodal information within the evidence page, thereby improving both the accuracy and robustness of generated answers. Our contributions are summarized as follows:

1. We propose FUMA, an end-to-end DocVQA framework that reconstructs evidence localization from a similarity-based retrieval task into a precise selection problem on unique semantics, fundamentally addressing the core challenge that semantically similar pages cannot be reliably distinguished in conventional representation spaces.

2. We theoretically prove the existence of a bijection between document pages and their question-conditioned unique semantics. Building on this foundation, we design a Flow-Based Unique Semantic Mapping module that learns an invertible transformation path by combining fine-grained cross-modal attention with flow-based likelihood regularization, ensuring that each page obtains an independent and identifiable unique semantic representation.

3. Extensive experiments conducted on multiple benchmark datasets demonstrate that FUMA significantly outperforms existing methods in both evidence page localization and answer generation quality, while the ablation studies further verify the effectiveness of the unique semantic modeling.

2 Related Work

2.1 Document Visual Question Answering

Document visual question answering (DocVQA) has evolved from traditional OCR-based pipelines to end-to-end generative frameworks. Early approaches relied on OCR for text extraction followed by BERT-based text understanding models (Mathew et al., 2021), which suffer from error propagation and limited visual understanding. The

LayoutLM family (Xu et al., 2020; Huang et al., 2022) substantially improves performance on structured documents via layout-aware pretraining that jointly encodes text, spatial positions, and visual features. More recent end-to-end methods, such as Donut (Kim et al., 2022) and Pix2Struct (Lee et al., 2023), remove explicit OCR dependence using vision–language Transformers, while multimodal large language models (mPLUG-PaperOwl (Hu et al., 2024) and UReader (Ye et al., 2023)) achieve strong zero-shot capability through instruction tuning. Interpretability has become a key research direction: DocVXQA (Souibgui et al., 2025) provides visual attention heatmaps, and (Li et al., 2025) generates multimodal explanations. Multi-page benchmarks such as MMVQA (Ding et al., 2024) and NiM-Benchmark (Thakkar et al., 2025) have further advanced comprehensive document intelligence, while provenance-based evaluation methods (Nourbakhsh et al., 2025) address hallucination risks in real-world deployment.

2.2 Flow-Based Model

Flow-based models map a simple distribution to a complex data distribution via invertible transformations, enabling explicit density estimation and exact likelihood computation. Early architectures include NICE (Dinh et al., 2014), which employs coupling layers; RealNVP (Dinh et al., 2017), which uses affine transformations; Glow (Kingma and Dhariwal, 2018), which introduces invertible 1×1 convolutions; and FFJORD (Grathwohl et al., 2018), which realizes continuous normalizing flows via ordinary differential equations (ODEs). More recently, Flow Matching (Lipman et al., 2023) systematizes probability flow learning as a simulation-free alternative to diffusion models. Subsequent extensions include Conditional Flow Matching (Tong et al., 2023) for conditional generation, Rectified Flow (Liu et al., 2022) for minimizing transport curvature, and Local Flow Matching (Xu et al., 2026) for stepwise factorization. Diff2Flow (Schusterbauer et al., 2025) facilitates knowledge transfer from diffusion models, while Reflected Flow Matching (Xie et al., 2024) addresses domain constraints. In structured generation, LayoutFlow (Guerreiro et al., 2024) applies flow matching to layout design and adopts a multimodal base distribution. FlowVQA (Singh et al., 2024) introduces a flowchart reasoning benchmark, highlighting spatial–logical understanding that is relevant to document VQA scenarios.

3 Method

This section systematically presents the proposed methodological framework. Section 3.1 first provides a theoretical analysis, demonstrating that **under the given problem conditions, there exists a bijective relationship between a document page and its unique semantics, and the answer is necessarily contained within the unique semantics of the evidence page**. Based on this insight, we introduce the Flow-based concept with an invertible structure and propose the **F**low-based Page **U**nique Semantic **M**apping **A**rchitecture (**FUMA**). The architecture consists of three main components: (1) a **Pre-trained Encoder**, which embeds document images and questions into a shared vector space; (2) the **Flow-based Unique Semantic Mapping** module (Section 3.2), which learns a bijection between each page and its unique semantics through multiple Flow-based blocks; and (3) a **Multi-expert Decoder** (Section 3.3), which performs complementary modeling and decoding of multimodal fine-grained information from the evidence page to generate precise answers. The overall architecture is illustrated in Figure 2.

Specifically, given an input document $D = \{I_j\}_{j=1}^{N}$ and a question Q, we first employ a text encoder f_{enc}^{t} and an image encoder f_{enc}^{i} to obtain the question vector q and page vectors i_j:

$$q = f_{enc}^{t}(Q), \quad i_j = f_{enc}^{i}(I_j),\ j = 1, \ldots, N \tag{1}$$

Then, the Flow-Based Unique Semantic Mapping $\mathcal{F}_U$ maps each page to its corresponding unique semantics u_j:

$$U = \mathcal{F}_U(\{i_j\}_{j=1}^{N}, q) = \{u_j\}_{j=1}^{N} \tag{2}$$

The evidence page I_k is identified by computing the similarity between q and each u_j. Based on the unique semantics u_k of the evidence page, the multi-expert decoder f_{dec} generates the final answer A:

$$A = f_{dec}(u_k, q) \tag{3}$$

Furthermore, Section 3.4 elaborates on the training objectives of FUMA, which aim to enhance both the localization ability for evidence pages and the accuracy of answer generation. In addition, flow-based likelihood regularization is employed to learn an approximately bijective mapping from page representations to their unique semantics.

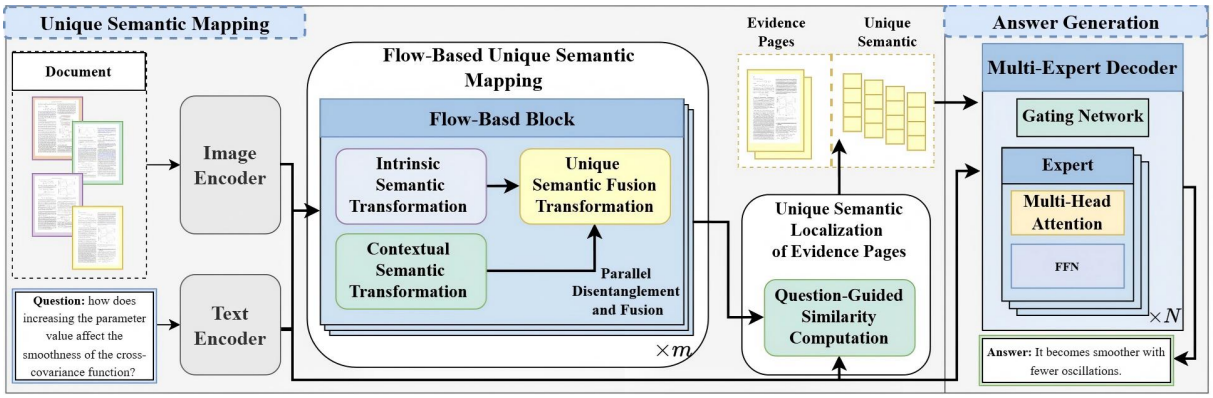

Figure 2: FUMA consists of three main components: (1) a pretrained image encoder and a pretrained text encoder, which embed the input document images and the natural-language question into vector representations; (2) a Flow-Based Unique Semantic Mapping module (introduced in 3.2), built by stacking Flow-Based Blocks, which learns a bijection from each page to its page-wise unique semantics; and (3) a multi-expert decoder (introduced in 3.3), which leverages a multi-expert collaboration mechanism to generate accurate answers conditioned on the page-wise unique semantics.

3.1 Theoretical Analysis

In DocVQA, the core challenge lies not only in the confusion caused by semantically similar pages but, more fundamentally, in the requirement that question-driven evidence localization depends on a separable semantic space. Specifically, under the given question conditions, each page should possess a unique and distinguishable semantic representation. In this way, different pages can be clearly differentiated at the semantic level, transforming evidence localization into the precise selection of unique semantics and providing a stable and traceable foundation for answer generation.

To motivate our architectural design and establish a structural foundation for the proposed framework, we formally analyze three key properties of the question-conditioned semantic space: (1) whether a unique semantic representation exists for each page under the given question; (2) whether the mapping from a page to its unique semantics is bijective, ensuring invertibility and lossless representation; and (3) the relationship between the question's answer and the unique semantics of the evidence page. Based on these analyses, we propose and prove the following propositions.

Proposition 1. *Given a question q and a document $D = \{i_j\}_{j=1}^{N}$, there exists a semantic space $\mathcal{S}$ and a mapping $\phi_q : D \to \mathcal{S}$ such that, for any two distinct pages $i_k, i_l \in D$ $(k \neq l)$, their semantic representations $u_k = \phi_q(i_k)$ and $u_l = \phi_q(i_l)$ satisfy $u_k \neq u_l$:*

Proposition 2. *Given a question q and a document $D = \{i_j\}_{j=1}^{N}$, let $\mathcal{U}_D = \{u_j\}_{j=1}^{N}$ denote the set of unique semantic representations of the pages, where $u_j = \phi_q(i_j)$. Then, the mapping*

$$\phi_q|_D : D \to \mathcal{U}_D, \quad i_j \mapsto u_j \tag{4}$$

is bijective.

Proposition 3. *Given a question q and a document $D = \{i_j\}_{j=1}^{N}$, suppose the answer a is located in the evidence page $i^* \in D$. Let $u_a = f_{enc}^{t}(a, q) \in \mathcal{S}$ denote the semantic representation of the answer, and $u_{i^*} = \phi_q(i^*)$ denote the unique semantics of the corresponding evidence page. Then,*

$$u_a \sqsubseteq u_{i^*} \tag{5}$$

where the containment relation is defined as:

$$u_a \sqsubseteq u_{i^*} \iff I(u_a; u_{i^*}) = H(u_a) \tag{6}$$

where $I(\cdot;\cdot)$ denotes the mutual information and $H(\cdot)$ denotes the Shannon entropy.

The proofs of the above propositions are provided in Appendix A.2. Collectively, these three propositions constitute the theoretical foundation of our framework: Proposition 1 ensures that, given a question q, each page possesses a unique and distinguishable semantic representation; Proposition 2 further guarantees the existence of an invertible bijective relationship between each page and its unique semantics, enabling lossless semantic mapping; and Proposition 3 demonstrates that the semantic representation of the answer is entirely contained within the unique semantics of the evidence page. It should be emphasized that these

propositions serve as *design principles* that justify the use of bijective, invertible transformations, rather than as formal guarantees about the representations learned by the neural network in practice. Specifically, they characterize the idealized conditions under which evidence localization reduces to a well-posed selection problem, and thereby guide our architectural design choices.

3.2 Flow-Based Unique Semantic Mapping

According to Proposition 1 and Proposition 2, there exists a bijective relationship between each document page and its unique semantics under the given question conditions. To realize this property within the model, we introduce the concept of a Flow-Based Model, which constructs the mapping relationship through a composition of invertible transformations. Specifically, we design a **Flow-Based Unique Semantic Mapping** $\mathcal{F}_U$, composed of m stacked invertible Flow-Based Blocks. These blocks progressively transform the page representations layer by layer, ultimately mapping each page i_j to its corresponding unique semantics u_j:

$$u_j = \mathcal{F}_U(\{i_j\}_{j=1}^N, q) = \mathcal{F}_m^u \circ \cdots \circ \mathcal{F}_1^u(\{i_j\}_{j=1}^N, q) \tag{7}$$

Each Flow-Based Block $\mathcal{F}_b^u$ adopts a dual-path architecture based on the principle of parallel disentanglement and fusion, which simultaneously models the intrinsic semantics and contextual semantics of a page to achieve a fine-grained characterization of its unique semantics. As illustrated in Figure 3, each block consists of three core components: (1) **Intrinsic Semantic Transformation** $\mathcal{S}_b^u$, which extracts the intrinsic semantic features s_j^b of the page under question guidance, focusing on question-relevant content; (2) **Contextual Semantic Transformation** $\mathcal{C}_b^u$, which captures the global semantics c_j^b induced by the interactions among document pages; and (3) **Unique Semantic Fusion Transformation** $\mathcal{T}_b^u$, which fuses the two semantic streams under the given question to generate a more discriminative unique semantic representation u_j^b. This design enables the model to jointly capture both local and global page properties, thereby enhancing the discriminability of the unique semantics.

For the b-th Flow-Based Block ($b \geq 1$), the input consists of the previous layer's unique semantic representation u_j^{b-1} and contextual semantic representation c_j^{b-1}. Under a question-guided multi-head attention mechanism, the block first performs the intrinsic and contextual transformations in parallel and subsequently integrates them through a cross-attention-based fusion. The mathematical formulations of the three core components are as follows.

Intrinsic Semantic Transformation $\mathcal{S}_b^u$:

$$s_j^b = \mathcal{S}_b^u(q, u_j^{b-1}) = \mathcal{F}_N(\mathcal{M}(q, u_j^{b-1}, u_j^{b-1})) \tag{8}$$

where $\mathcal{M}(\cdot)$ denotes the multi-head attention mechanism and $\mathcal{F}_N(\cdot)$ denotes a feed-forward neural network. The detailed operations are described in Appendix A.1. This transformation adaptively reweights the page semantics under the guidance of question q, ensuring that s_j^b focuses on question-relevant features while filtering out irrelevant or redundant information.

Contextual Semantic Transformation $\mathcal{C}_b^u$:

$$c_j^b = \mathcal{C}_b^u(q, c_j^{b-1}) = \mathcal{F}_N(\mathcal{M}(q, c_j^{b-1}, c_j^{b-1})) \tag{9}$$

This transformation models the relative positional and semantic dependencies among document pages from a global perspective, enabling c_j^b to encode the contextual semantics of each page within the document.

Unique Semantic Fusion Transformation $\mathcal{T}_b^u$:

$$u_j^b = \mathcal{T}_b^u(q, c_j^b, s_j^b) = \mathcal{F}_N(\mathcal{M}(q, c_j^b, s_j^b)) \tag{10}$$

This fusion module integrates the intrinsic and contextual semantics through a question-guided cross-attention mechanism, producing the updated unique semantics u_j^b.

The entire Flow-Based Unique Semantic Mapping module $\mathcal{F}_U$ consists of m stacked Flow-Based Blocks, each following a progressive parallel disentanglement–collaborative fusion pattern. As a result, the final output u_j^m exhibits both uniqueness and discriminability under the given question, providing a robust representation for evidence page localization.

At the input stage ($b = 1$), a relative representation-based initialization strategy is applied. First, the mean embedding of all page representations is used as the global contextual initialization:

$$c_j^0 = \frac{1}{N} \sum_{k=1}^{N} i_k, \quad \forall j \in \{1, \ldots, N\} \tag{11}$$

which captures the document's global semantic baseline. Then, the unique semantics are initialized by computing the deviation of each page from

the global context:

$$u_j^0 = i_j - c_j^0 = i_j - \frac{1}{N}\sum_{k=1}^{N} i_k \tag{12}$$

thereby achieving zero-centered representations that facilitate stable bijective learning and invertible transformations.

Through the layered semantic transformations, $\mathcal{F}_U$ maps the page representations into a question-conditioned unique semantic space, where $\{u_j\}_{j=1}^N$ fully capture inter-page distinctions and semantic discriminability.

Unique Semantic Localization. After obtaining the unique semantic representations of all pages $U = \{u_j\}_{j=1}^N$, and according to Proposition 3, the semantics of the answer are contained within the unique semantics of the evidence page. To this end, evidence localization is achieved by computing the similarity between the question representation and the unique semantics of each page:

$$z_j = \mathrm{sim}(q, u_j), \quad k = \arg\max_j z_j \tag{13}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity function.

As indicated by Proposition 2, the evidence page I_k and its corresponding unique semantics u_k are bijectively related, ensuring information preservation. Consequently, this semantic-based localization mechanism guarantees the traceability of the identified evidence and provides a semantically complete conditional representation for answer generation.

3.3 Multi-Expert Decoder

To fully exploit the multimodal fine-grained information embedded in the evidence page, we design a **Multi-Expert Decoder** that generates answers by complementarily modeling the unique semantics of the evidence page from multiple expert perspectives.

The decoder consists of N experts, denoted as $\mathcal{E} = \{E_1, \ldots, E_N\}$, where each expert is implemented using a multi-head attention mechanism followed by a feed-forward neural network:

$$h_i = E_i(q, u_k) = \mathcal{F}_N(\mathcal{M}(q, u_k, u_k)) \tag{14}$$

To improve computational efficiency and focus on the most relevant experts, a Top-K selection mechanism is introduced. A gating network

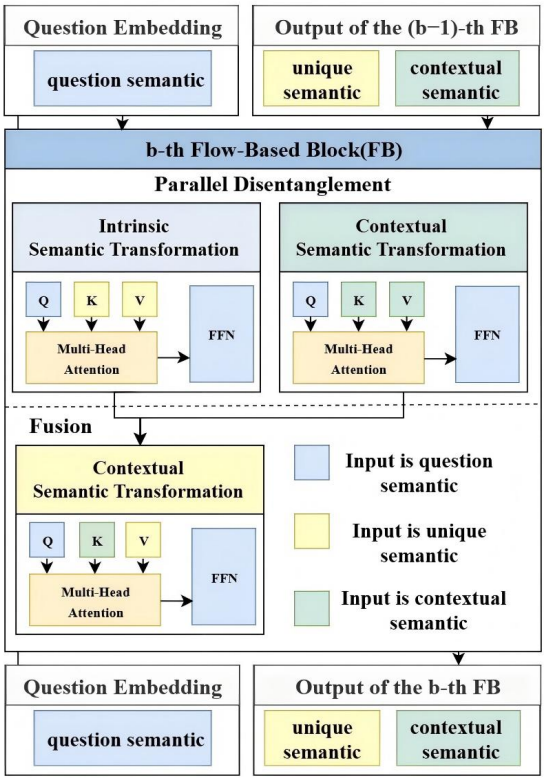

Figure 3: Schematic illustration of the parallel–fusion architecture of the Flow-Based Block. Each block contains two parallel transformation modules—Intrinsic Semantic Transformation and Contextual Semantic Transformation that extract intrinsic semantics and contextual semantics, respectively. These two semantic streams are then integrated through the Unique Semantic Fusion Transformation, producing the updated unique semantics. All transformations are guided by the question

$g(q, u_k)$ computes the weights of all experts, after which the Top-K experts are selected and their outputs are aggregated in a weighted manner to generate the final answer:

$$A = \sum_{j=1}^{K} \tilde{w}_{i_j} \cdot h_{i_j} \tag{15}$$

where $\tilde{w}_{i_j}$ denotes the normalized expert weights. This mechanism enables the model to dynamically select and integrate the most relevant expert knowledge based on the characteristics of the question and the evidence page, thereby enhancing answer accuracy and robustness.

3.4 Training Objective

Given a training sample (D, Q, A, k^*), where $D = \{I_j\}_{j=1}^N$ denotes a document containing N pages, Q is the question, $A = (a_1, \ldots, a_T)$ represents the target answer sequence, and k^* is the index

Table 1: Performance comparison on four benchmark datasets. The best results are marked in **bold**.

Method	NiM-Benchmark			MMlongBench-Doc			DUDE			DocVQA		
	EM	F1	ANLS	EM	F1	ANLS	EM	F1	ANLS	EM	F1	ANLS
Hi-VT5 (Tito et al., 2023)	63.71	79.68	82.14	61.12	72.26	65.17	60.97	69.58	73.51	49.87	73.20	76.50
LayoutLMv3 (Huang et al., 2022)	81.55	88.64	92.57	87.52	89.63	92.89	71.54	79.58	83.41	54.68	84.92	86.71
DocFormer (Appalaraju et al., 2021)	74.63	84.98	87.43	77.45	82.58	84.23	68.24	74.35	76.77	55.23	83.97	87.84
GRAM (Blau et al., 2024)	74.87	85.69	88.11	79.46	82.35	83.97	68.57	77.52	79.67	53.97	83.74	85.96
MP-FIRE (Yu et al., 2025)	78.59	87.52	89.58	85.48	89.14	90.74	69.52	78.43	80.01	51.69	79.63	82.20
FUMA (Ours)	**84.44**	**94.35**	**96.72**	**89.11**	**94.42**	**95.03**	**73.72**	**87.86**	**91.13**	**56.44**	**84.34**	**88.96**

of the evidence page containing the answer. The proposed FUMA is trained in an end-to-end manner by jointly optimizing three objectives: (i) evidence page localization; (ii) answer generation; and (iii) bijective regularization.

Evidence Localization Loss As described in Section 3.2, the matching score for each page is computed as $z_j = \mathrm{sim}(q, u_j)$, and then normalized using a softmax function to form a probability distribution:

$$P(j \mid D, Q) = \frac{\exp(z_j)}{\sum_{l=1}^{N} \exp(z_l)} \tag{16}$$

The localization loss adopts a cross-entropy form:

$$\mathcal{L}_{\text{loc}} = -\log P(k^* \mid D, Q) \tag{17}$$

which encourages the model to assign the highest probability to the true evidence page.

Answer Generation Loss The Multi-Expert Decoder generates the answer conditioned on (u_{k^*}, q) using a teacher-forcing strategy:

$$\mathcal{L}_{\text{ans}} = -\sum_{t=1}^{T} \log P(a_t \mid a_{<t}, u_{k^*}, q) \tag{18}$$

directly optimizing the model's ability to generate accurate answers based on the unique semantics of the evidence page.

Bijective Regularization Loss. To ensure that $\mathcal{F}_U$ learns an approximately bijective mapping, we introduce a flow-based regularization term (Rezende and Mohamed, 2015; Dinh et al., 2017). The negative log-likelihood of the unique semantics under the base distribution p_U can be expressed as:

$$\begin{aligned}\mathcal{L}_{\text{flow}} &= -\mathbb{E}_{i_j \sim p_I}\left[\log p_U\big(\mathcal{F}_U(i_j, q)\big)\right] \\ &= -\mathbb{E}_{i_j \sim p_I}[\log p_I(i_j) + \log|\det J_{\mathcal{F}_U}(i_j, q)|]\end{aligned} \tag{19}$$

where p_I denotes the data distribution of page representations and $J_{\mathcal{F}_U}(i_j, q) = \frac{\partial \mathcal{F}_U(i_j, q)}{\partial i_j}$ is the Jacobian matrix of the transformation. Assuming a standard Gaussian base distribution $p_U = \mathcal{N}(\mathbf{0}, \mathbf{I})$, and noting that $\log p_I(i_j)$ is constant with respect to the model parameters, the practical regularization term becomes:

$$\begin{aligned}\mathcal{L}_{\text{flow}} = \mathbb{E}_{i_j \sim p_I}\Bigg[&\frac{1}{2}\|\mathcal{F}_U(i_j, q)\|_2^2 \\ &- \log|\det J_{\mathcal{F}_U}(i_j, q)|\Bigg]\end{aligned} \tag{20}$$

where the first term constrains the compactness of the semantic space, and the second term promotes local invertibility of the transformation. Together, they ensure that different pages are mapped to distinguishable unique semantics while preserving information.

Joint Objective. The overall FUMA training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + \mathcal{L}_{\text{loc}} + \gamma\, \mathcal{L}_{\text{flow}} \tag{21}$$

where $\gamma > 0$ is a balancing hyperparameter. This joint objective enables the model to simultaneously learn evidence localization, answer generation, and invertible mapping preservation, thereby achieving precise evidence identification and accurate answer generation across multi-page documents.

4 Experiments

4.1 Experimental Setup

We evaluate our proposed method on four benchmark datasets: **(1) NiM-Benchmark** (Thakkar et al., 2025), **(2) DUDE** (Landeghem et al., 2023), **(3) MMlongBench-Doc** (Ma et al., 2024), and **(4) DocVQA** (Mathew et al., 2021). We compare FUMA with the following state-of-the-art methods: **Hi-VT5** (Tito et al., 2023), **LayoutLMv3** (Huang et al., 2022), **DocFormer** (Appalaraju et al., 2021), **GRAM** (Blau et al., 2024), and **MP-FIRE** (Yu et al., 2025). We adopt three standard evaluation metrics to assess model performance: **Exact Match (EM)**, **F1 Score**, and **Average Normalized Levenshtein Similarity (ANLS)**. Further experimental details are provided in Appendix A.3.

Table 2: Ablation results on NiM-Benchmark, MMlongBench-Doc, DUDE, and DocVQA.

Variant	NiM-Benchmark			MMlongBench-Doc			DUDE			DocVQA		
	EM	F1	ANLS	EM	F1	ANLS	EM	F1	ANLS	EM	F1	ANLS
w/o Intrinsic Semantic Transform	79.82	89.54	92.33	81.45	88.26	89.17	68.91	82.47	85.68	53.15	80.26	84.12
w/o Context-Semantic Transformation	77.48	87.30	90.15	78.32	85.43	87.24	66.73	79.85	83.28	52.34	78.67	82.58
w/o Unique-semantic Fusion Transform	69.15	78.63	82.74	65.42	74.58	77.35	58.26	69.72	74.86	48.32	72.18	76.92
w/o Multi-Expert Collaboration	74.26	83.92	87.58	73.58	81.24	83.67	63.45	75.38	79.42	50.87	75.94	80.15
w/o Flow-Based Unique Semantic Mapping	61.84	71.26	75.43	57.36	68.94	71.58	52.47	63.15	68.42	44.18	66.02	70.35
FUMA (Ours)	**84.44**	**94.35**	**96.72**	**89.11**	**94.42**	**95.03**	**73.72**	**87.86**	**91.13**	**56.44**	**84.34**	**88.96**

Table 3: Evidence-page localization accuracy (%) on multi-page document datasets.

Method	DUDE	MMlongBench-Doc
Hi-VT5	65.32	58.91
LayoutLMv3	74.58	70.23
DocFormer	71.45	68.77
GRAM	76.89	72.14
MP-FIRE	82.71	77.25
FUMA (Ours)	**91.24**	**88.67**

4.2 Main Results

Table 1 presents the performance comparison across four benchmark datasets. FUMA achieves the best results on all datasets and evaluation metrics, demonstrating its overall superiority and robustness. On the NiM-Benchmark dataset, FUMA surpasses LayoutLMv3 by 2.89%, 5.71%, and 4.15% in EM, F1, and ANLS, respectively, verifying the effectiveness of the proposed unique semantic mapping in handling diverse document types. On the long-document benchmark MMlongBench-Doc, FUMA outperforms MP-FIRE by 3.63% (EM), 5.28% (F1), and 4.29% (ANLS), where the larger margin highlights its advantage in modeling complex multi-page semantics. The consistent superiority on the DUDE dataset further indicates that the flow-based bijective mapping effectively maintains fine-grained discriminative capability across documents with complex layouts. Moreover, the stable performance on DocVQA confirms FUMA's strong generalization ability, showing that the framework preserves single-page understanding while enhancing multi-page reasoning performance.

4.3 Ablation Study

To evaluate the effectiveness of each component, we conduct ablation experiments as shown in Table 2. Removing the Flow-Based Unique Semantic Mapping leads to the most significant performance drop (a 25.48% decrease in F1 on MMlongBench-Doc), indicating that the bijective semantic mapping serves as the core foundation of the proposed framework. Eliminating the unique semantic fusion, contextual semantic transformation, or intrinsic semantic modeling modules all result in consistent performance degradation, verifying their complementary roles in fine-grained feature representation. Additionally, removing the multi-expert collaboration module causes an average F1 drop of 10.43%, demonstrating that the collaborative expert mechanism is crucial for enhancing multimodal discriminative capability.

4.4 Evidence-Page Localization Analysis

Table 3 reports the evidence page localization accuracy on the DUDE and MMlongBench-Doc datasets. FUMA achieves localization accuracies of 91.24% and 88.67% on DUDE and MMlongBench-Doc, respectively, outperforming the strongest baseline, MP-FIRE, by 8.53% and 11.42%. These substantial improvements clearly demonstrate the superiority of our approach over existing methods. More importantly, this gain directly validates the core contribution of this work: by establishing a bijective mapping from document pages to question-conditioned unique semantics, FUMA transforms evidence localization from a confusion-prone similarity retrieval problem into an accurate selection process within a well-separated semantic space. This bijective semantic mapping effectively resolves the key challenge of distinguishing semantically similar pages, providing a stable and reliable evidence foundation for subsequent answer generation.

4.5 Backbone Generalizability Analysis

To assess the backbone-agnostic property of FUMA, we evaluate three encoder pairings of increasing capacity: ViT-Base (Dosovitskiy et al., 2021) + BERT-Base (Devlin et al., 2019), Swin-Base (Liu et al., 2021) + RoBERTa-Large (Liu et al., 2020), and Florence-2 (Xiao et al., 2024) + Qwen2.5 (Qwen et al., 2025), comparing against MP-FIRE under identical settings on DUDE (Ta-

Table 4: Comparison with representative MLLMs on the DUDE dataset. AC denotes answer-conditioned page localization accuracy. Best results are in **bold**.

Method	Size	EM	F1	ANLS	AC
Qwen2.5-VL (Bai et al., 2025)	7B	65.89	77.84	81.27	52.67
InternVL2.5 (Chen et al., 2025)	8B	63.56	76.72	82.13	54.13
LLaVA-1.6 (Liu et al., 2024)	13B	66.43	80.57	82.34	60.57
FUMA (Ours)	**4B**	**73.72**	**87.86**	**91.13**	**91.24**

Table 5: Backbone generalizability comparison between FUMA and MP-FIRE on the DUDE dataset across three encoder configurations. Best results per configuration are in **bold**.

Method	Vision Encoder	Text Encoder	EM	F1	ANLS	AC
MP-FIRE	ViT-Base	BERT-Base	64.21	79.02	83.24	81.84
FUMA	ViT-Base	BERT-Base	**70.54**	**83.27**	**88.62**	**86.43**
MP-FIRE	Swin-Base	RoBERTa-Large	68.98	84.44	86.01	87.32
FUMA	Swin-Base	RoBERTa-Large	**72.54**	**87.27**	**90.83**	**88.43**
MP-FIRE	Florence-2	Qwen2.5	69.11	83.02	86.12	87.07
FUMA	Florence-2	Qwen2.5	**72.89**	**87.01**	**90.34**	**90.62**

ble 5). FUMA consistently outperforms MP-FIRE across all configurations on every metric, with stable EM gains of $+6.33/+3.56/+3.78$ and AC gains of $+4.59/+1.11/+3.55$ from lightweight to large-scale. The uniformity of these margins across architectures of substantially different capacity confirms that the improvements stem from the proposed semantic mapping mechanism rather than backbone expressiveness.

4.6 Comparison with Multimodal Large Language Models

While MLLMs have demonstrated strong generative and zero-shot capabilities on document understanding benchmarks (Bai et al., 2025; Chen et al., 2025; Liu et al., 2024), Table 4 shows that all evaluated MLLMs exhibit substantially lower evidence page localization accuracy (AC) despite competitive EM/F1 scores, suggesting that they often generate plausible answers without precisely grounding semantics to the correct evidence page.

In contrast, FUMA surpasses all competing multimodal large language models across every evaluation metric, achieving a substantial improvement of 30.67 percentage points in localization accuracy (AC) over the strongest baseline. This improvement is attributable to the proposed query-conditioned semantic mapping mechanism, further underscoring the critical role of semantic bijectivity in multi-page document visual question answering (DocVQA).

5 Conclusion

This paper addresses the fundamental challenge of evidence-page localization in multi-page DocVQA, where semantic similarity across pages causes confusion in conventional similarity-based retrieval approaches. We theoretically prove, via three formal propositions, that each document page admits a unique question-conditioned semantic representation and that a bijective mapping exists between pages and their unique semantics. Building on this foundation, we propose FUMA, which reformulates evidence localization as precise selection over a well-separated unique semantic space, integrating a flow-based invertible transformation module with a multi-expert collaborative decoder. Extensive experiments on four benchmarks demonstrate consistent superiority in both localization accuracy and answer generation quality. Comparisons across diverse backbone configurations and against representative MLLMs further confirm that the gains stem from the proposed semantic mapping mechanism rather than model scale, underscoring the broader applicability of bijective semantic learning to multi-page document understanding.

Limitations

Despite strong empirical performance, FUMA has several limitations. First, the flow-based invertible mapping introduces non-trivial training complexity: the Jacobian log-determinant computation, though tractable via chain rule decomposition, still incurs additional computational overhead. Second, FUMA relies on evidence-level page supervision during training, limiting its applicability in weakly supervised or annotation-scarce settings. Third, the localization mechanism operates at the page level and does not support sub-page fine-grained evidence grounding; for documents where the answer spans fine-grained regions such as table cells or chart elements, the model may generate incomplete answers. Future work will explore more stable invertible learning objectives, weakly supervised localization strategies, and hierarchical grounding mechanisms to improve robustness and scalability to long-document scenarios.

Acknowledgements

This research was funded by the National Natural Science Foundation of China (No.52538002)

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 993–1003.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Tsachi Blau, Sharon Fogel, Roi Ronen, Alona Golts, Shahar Tsiper, Elad Ben Avraham, Aviad Aberdam, Roy Ganz, and Ron Litman. 2024. GRAM: Global Reasoning for Multi-Page VQA . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15598–15607, Los Alamitos, CA, USA. IEEE Computer Society.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2025. M3docvqa: Multi-modal multi-page multi-document understanding. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 6237–6247, Los Alamitos, CA, USA. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Yihao Ding, Zhe Huang, Runlin Wang, YanHang Zhang, Xianru Chen, Yuzhong Ma, Hyunsuk Chung, and Soyeon Caren Han. 2022. V-doc: Visual questions answers with documents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21492–21498.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024. Mmvqa: A comprehensive dataset for investigating multipage multimodal information retrieval in pdf-based visual question answering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6243–6251. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Laurent Dinh, David Krueger, and Yoshua Bengio. 2014. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2017. Density estimation using real nvp. *Preprint*, arXiv:1605.08803.

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. 2018. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *Preprint*, arXiv:1810.01367.

Julian Jorge Andrade Guerreiro, Naoto Inoue, Kento Masui, Mayu Otani, and Hideki Nakayama. 2024. Layoutflow: Flow matching for layout generation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXVI*, page 56–72, Berlin, Heidelberg. Springer-Verlag.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-paperowl: Scientific diagram analysis with the multimodal large language model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 6929–6938, New York, NY, USA. Association for Computing Machinery.

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9989–9999.

Xin Huang and Kye Min Tan. 2025. Beyond text: Unlocking true multimodal, end-to-end rag with tomoro colqwen3.

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*, MM '22, page 4083–4091, New York, NY, USA. Association for Computing Machinery.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5648–5656.

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*, page 498–517, Berlin, Heidelberg. Springer-Verlag.

Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jordy Van Landeghem, Rafał Powalski, Rubèn Tito, Dawid Jurkiewicz, Matthew Blaschko, Łukasz Borchmann, Mickaël Coustaty, Sien Moens, Michał Pietruszka, Bertrand Ackaert, Tomasz Stanisławek, Paweł Józiak, and Ernest Valveny. 2023. Document understanding dataset and evaluation (dude). In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19471–19483.

Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. Formnet: Structural encoding beyond sequential modeling in form document information extraction. *Preprint*, arXiv:2203.08411.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Kun Li, George Vosselman, and Michael Ying Yang. 2025. Multimodal rationales for explainable visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 191–201.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *ArXiv*, abs/2209.03003.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Roberta: A robustly optimized bert pretraining approach.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *Preprint*, arXiv:2103.14030.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. MMLONGBENCH-DOC: Benchmarking long-context document understanding with visualizations. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208.

Armineh Nourbakhsh, Siddharth Parekh, Pranav Shetty, Zhao Jin, Sameena Shah, and Carolyn Rose. 2025. Where is this coming from? making groundedness count in the evaluation of document VQA models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5341–5361, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinhui Pang, Xinyun Yang, Xiaoyao Qiu, Zixuan Wang, and Taisheng Huang. 2024. Mmaf: Masked multimodal attention fusion to reduce bias of visual features for named entity recognition. *DATA INTELLIGENCE*, 6(4):1114–1133.

Yang Qin, Huiming Xie, Yujie Li, Benying Tan, and Shuxue Ding. 2025. Enhancing intermodal interaction for unified vision-language understanding and generation. *DATA INTELLIGENCE*, 7(2):358–380.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Danilo Jimenez Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, page 1530–1538. JMLR.org.

Johannes Schusterbauer, Ming Gui, Frank Fundel, and Björn Ommer. 2025. Diff2flow: Training flow matching models via diffusion model alignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28347–28357.

Shubhankar Singh, Purvi Chaurasia, Yerram Varun, Pranshu Pandya, Vatsal Gupta, Vivek Gupta, and Dan Roth. 2024. FlowVQA: Mapping multimodal logic in visual question answering with flowcharts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1330–1350, Bangkok, Thailand. Association for Computational Linguistics.

Mohamed Ali Souibgui, Changkyu Choi, Andrey Barsky, Kangsoo Jung, Ernest Valveny, and Dimosthenis Karatzas. 2025. DocVXQA: Context-aware visual explanations for document question answering. In *Forty-second International Conference on Machine Learning*.

Parth Thakkar, Ankush Agarwal, Prasad Kasu, Pulkit Bansal, and Chaitanya Devaguptapu. 2025. Finding needles in images: Can multi-modal llms locate fine details? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23626–23648.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Fatras Kilian, and Guy Wolf. 2023. Conditional flow matching: Simulation-free dynamic optimal transport.

Yiwen Wang, Xiaobing Zhao, Xiaoke Qi, Bo Chen, Chuanlian Ma, and Yang Xu. 2025. A large language model evaluation method for legal case retrieval. *DATA INTELLIGENCE*, 7(2):440–460.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4818–4829.

Tianyu Xie, Yu Zhu, Longlin Yu, Tong Yang, Ziheng Cheng, Shiyue Zhang, Xiangyu Zhang, and Cheng Zhang. 2024. Reflected flow matching. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Chen Xu, Xiuyuan Cheng, and Yao Xie. 2026. Local flow matching generative models. *IEEE Transactions on Information Theory*, 72(5):3339–3358.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1192–1200, New York, NY, USA. Association for Computing Machinery.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Fei Huang. 2023. UReader: Universal OCR-free visually-situated language understanding with multimodal large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2841–2858, Singapore. Association for Computational Linguistics.

Yongqi Yu, Jinxu Zhang, and Yu Zhang. 2025. Mp-fire: An end-to-end cross-modal framework for complex multi-page document question answering. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

A Appendix

A.1 Notation and Supplementary Knowledge

A.1.1 Notation Table

The notations used in this paper are shown in Table 6.

A.1.2 Multi-Head Attention Mechanism

The multi-head attention mechanism $\mathcal{M}(\cdot)$ is a core component in our Flow-Based Blocks and Multi-Expert Decoder. Given a query Q, key K, and value V, the multi-head attention operation is defined as follows.

Single-Head Attention For a single attention head, the scaled dot-product attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{22}$$

where d_k is the dimension of the key vectors, and the scaling factor $\frac{1}{\sqrt{d_k}}$ prevents the dot products from growing too large in magnitude.

Multi-Head Attention To capture information from different representation subspaces, the multi-head attention mechanism employs H parallel attention heads. Each head h has its own learned linear projections:

$$\text{head}_h = \text{Attention}(QW_h^Q, KW_h^K, VW_h^V) \tag{23}$$

where $W_h^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_h^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, and $W_h^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ are the projection matrices for the h-th head.

The outputs of all heads are concatenated and linearly projected:

$$\mathcal{M}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)W^O \tag{24}$$

where $W^O \in \mathbb{R}^{Hd_v \times d_{\text{model}}}$ is the output projection matrix.

Question-Guided Attention In our framework, the attention mechanism is guided by the question q. For example, in the Intrinsic Semantic Transformation:

$$\mathcal{M}(q, u_j^{b-1}, u_j^{b-1}) \tag{25}$$

the question representation q serves as the query, while the previous layer's unique semantics u_j^{b-1} serves as both key and value. This allows the model to selectively attend to question-relevant features in the page representation.

Similarly, in the Unique Semantic Fusion Transformation:

$$\mathcal{M}(q, s_j^b, c_j^b) \tag{26}$$

the question q acts as the query, while the intrinsic semantics s_j^b and contextual semantics c_j^b are used as keys and values, enabling cross-attention between the two semantic streams.

A.1.3 Feed-Forward Neural Network

The feed-forward neural network $\mathcal{F}_N(\cdot)$ is applied to each position separately and identically. It consists of two linear transformations with a ReLU activation in between:

$$\mathcal{F}_N(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{27}$$

where $W_1 \in \mathbb{R}^{d_{\text{model}} \times d_{ff}}$, $b_1 \in \mathbb{R}^{d_{ff}}$, $W_2 \in \mathbb{R}^{d_{ff} \times d_{\text{model}}}$, and $b_2 \in \mathbb{R}^{d_{\text{model}}}$ are learnable parameters.

Residual Connection and Layer Normalization Both the multi-head attention and feed-forward network are wrapped with residual connections and layer normalization:

$$x' = \text{LayerNorm}(x + \mathcal{M}(\cdot)) \tag{28}$$

$$x'' = \text{LayerNorm}(x' + \mathcal{F}_N(x')) \tag{29}$$

This design is used to control training stability and gradient flow through deep networks.

A.1.4 Shannon Entropy and Mutual Information

Shannon entropy is a fundamental measure of uncertainty or information content in information theory. For a discrete random variable X with probability mass function $p(x)$, the Shannon entropy is defined as:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{30}$$

Conditional Entropy The conditional entropy $H(X|Y)$ measures the expected uncertainty in X given knowledge of Y:

$$H(X|Y) = -\sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log p(x|y) \tag{31}$$

This can also be written as:

$$H(X|Y) = H(X, Y) - H(Y) \tag{32}$$

where $H(X, Y)$ is the joint entropy of X and Y.

Mutual Information Mutual information (MI) measures the amount of information shared between two random variables, quantifying the reduction in uncertainty about one variable given knowledge of the other. For random variables X and Y, the mutual information is defined as:

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad (33)$$

where $p(x,y)$ is the joint probability distribution, and $p(x)$, $p(y)$ are the marginal distributions. Mutual information can be equivalently expressed in terms of entropy:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) && (34)\\ &= H(Y) - H(Y|X) && (35)\\ &= H(X) + H(Y) - H(X,Y) && (36)\end{aligned}$$

A.1.5 Cosine Similarity

The cosine similarity function $\mathrm{sim}(\cdot,\cdot)$ measures the cosine of the angle between two non-zero vectors. For vectors $\mathbf{a}$ and $\mathbf{b}$, it is defined as:

$$\mathrm{sim}(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|_2\|\mathbf{b}\|_2} = \frac{\sum_{i=1}^{d} a_i b_i}{\sqrt{\sum_{i=1}^{d} a_i^2}\sqrt{\sum_{i=1}^{d} b_i^2}} \quad (37)$$

where d is the dimension of the vectors. The cosine similarity ranges from -1 to 1, with higher values indicating greater similarity.

In the evidence localization stage (Section 3.2), we compute:

$$z_j = \mathrm{sim}(q, u_j) \quad (38)$$

to measure the relevance between the question representation q and each page's unique semantics u_j.

A.1.6 Gating Network for Expert Selection

The gating network $g(q,u_k)$ in the Multi-Expert Collaborative Decoder computes unnormalized importance scores for all experts. It is implemented as a simple feed-forward network:

$$g_i = g(q,u_k)_i = \mathbf{w}_i^\top \tanh(W_g[q;u_k] + b_g) \quad (39)$$

where $[q;u_k]$ denotes the concatenation of question and evidence page representations, $W_g \in \mathbb{R}^{d_g\times 2d_{\text{model}}}$ and $b_g \in \mathbb{R}^{d_g}$ are shared parameters, and $\mathbf{w}_i \in \mathbb{R}^{d_g}$ is the expert-specific weight vector.

Top-K Selection and Normalization After computing scores for all N experts, we select the Top-K experts with the highest scores:

$$\{i_1,\ldots,i_K\} = \text{TopK}(\{g_1,\ldots,g_N\}, K) \quad (40)$$

The selected expert weights are then normalized using softmax:

$$\tilde{w}_{i_j} = \frac{\exp(g_{i_j})}{\sum_{l=1}^{K}\exp(g_{i_l})}, \quad j = 1,\ldots,K \quad (41)$$

Finally, the weighted aggregation produces the answer representation:

$$A = \sum_{j=1}^{K} \tilde{w}_{i_j}\cdot h_{i_j} \quad (42)$$

where $h_{i_j} = E_{i_j}(q,u_k)$ is the output of the i_j-th expert.

A.1.7 Jacobian Determinant Computation

The Jacobian determinant $\det J_{\mathcal{F}_U}(i_j,q)$ in the bijective regularization loss measures the local volume change induced by the transformation $\mathcal{F}_U$.

Jacobian Matrix For a transformation $\mathbf{y} = \mathcal{F}_U(\mathbf{x},q)$ where $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, the Jacobian matrix is:

$$J_{\mathcal{F}_U}(\mathbf{x},q) = \frac{\partial\mathcal{F}_U(\mathbf{x},q)}{\partial\mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_d}{\partial x_1} & \cdots & \frac{\partial y_d}{\partial x_d} \end{bmatrix} \quad (43)$$

Log-Determinant Computing the determinant directly for high-dimensional matrices is computationally expensive. We leverage the chain rule for composed transformations:

$$\log|\det J_{\mathcal{F}_U}| = \sum_{b=1}^{m}\log|\det J_{\mathcal{F}_b^u}| \quad (44)$$

We adopt architectural constraints that ensure tractable Jacobian computation while maintaining expressiveness. To avoid numerical instability, we work directly with the log-determinant:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{i_j\sim p_I}\left[\frac{1}{2}\|\mathcal{F}_U(i_j,q)\|_2^2 - \log|\det J_{\mathcal{F}_U}(i_j,q)|\right] \quad (45)$$

The first term encourages the transformed representations to have small norms (fitting a standard Gaussian), while the second term prevents the transformation from collapsing dimensions.

Table 6: Notation Table

Notation	Description
N	Number of document pages (also number of experts in decoder)
i_j	Representation of the j-th page
I_j	The j-th page of the document
I_k	Evidence page
q	Question representation
u_j	Unique semantics of the j-th page
u_k	Unique semantics of the evidence page
u_j^b	Unique semantics of the j-th page at the b-th block
u_j^0	Initial unique semantics of the j-th page
s_j^b	Intrinsic semantics of the j-th page at the b-th block
c_j^b	Contextual semantics of the j-th page at the b-th block
c_j^0	Initial contextual semantics of the j-th page
U	Set of unique semantics of all pages, $U = \{u_j\}_{j=1}^N$
$\mathcal{F}_U$	Flow-Based Unique Semantic Mapping
$\mathcal{F}_b^u$	The b-th Flow-Based Block
m	Number of stacked Flow-Based Blocks
$\mathcal{S}_b^u$	Intrinsic Semantic Transformation at the b-th block
$\mathcal{C}_b^u$	Contextual Semantic Transformation at the b-th block
$\mathcal{T}_b^u$	Unique Semantic Fusion Transformation at the b-th block
$\mathcal{M}(\cdot)$	Multi-head attention mechanism
$\mathcal{F}_N(\cdot)$	Feed-forward neural network
$\mathrm{sim}(\cdot,\cdot)$	Cosine similarity function
z_j	Similarity score between question and unique semantics of page j
k	Index of the evidence page
$\mathcal{E}$	Set of experts in the decoder, $\mathcal{E} = \{E_1, \ldots, E_N\}$
E_i	The i-th expert in the decoder
h_i	Output of the i-th expert
h_{i_j}	Output of the i_j-th selected expert (Top-K)
K	Number of top experts selected
$g(q, u_k)$	Gating network for computing expert weights
$\tilde{w}_{i_j}$	Normalized weight of the i_j-th selected expert
A	Generated answer

A.2 Proof

A.2.1 Proof of Proposition 1

Proof. We define the semantic space $\mathcal{S}$ as the set of all possible joint representations of a page and the question. For a fixed question q and any page $i \in D$, define the mapping $\phi_q : D \to \mathcal{S}$ as

$$\phi_q(i) = \mathrm{Encode}(i, q) \tag{46}$$

where $\mathrm{Encode}(\cdot, \cdot)$ is a deterministic encoding function that jointly encodes the page and the question. It suffices to show that ϕ_q assigns distinct semantic representations to distinct pages.

We proceed by contradiction. Suppose there exist two distinct pages $i_k, i_l \in D$ ($k \neq l$) such that their (question-conditioned) semantics coincide:

$$u_k = \phi_q(i_k) = \phi_q(i_l) = u_l \tag{47}$$

Since $i_k \neq i_l$, at least one of the following cases must hold:

- **Case 1:** the page content differs. In particular, there exists a content element c such that $c \in i_k$ but $c \notin i_l$;
- **Case 2:** the page layout or structure differs;
- **Case 3:** the position of the page in the document differs.

Consider **Case 1**. Given the question q, the element c acts as a distinguishing signal. Define the indicator function

$$\mathbb{I}_c(i) = \begin{cases} 1, & \text{if } c \in i \\ 0, & \text{otherwise} \end{cases} \tag{48}$$

Then $\mathbb{I}_c(i_k) \neq \mathbb{I}_c(i_l)$. Under our basic assumption that $\mathrm{Encode}(i, q)$ is information-preserving with respect to page-specific signals relevant under q, it cannot map two pages that differ on such a distinguishing signal to the same representation. Hence,

$$\phi_q(i_k) \neq \phi_q(i_l) \tag{49}$$

which contradicts the assumption $u_k = u_l$.

For **Cases 2 and 3**, an analogous argument applies: differences in layout/structure or in document position also constitute distinguishing signals that must be reflected by an information-preserving joint encoder, again implying $\phi_q(i_k) \neq \phi_q(i_l)$. Therefore, the assumption $u_k = u_l$ is false, and we conclude that for any $i_k \neq i_l$, it holds that $u_k \neq u_l$. □

A.2.2 Proof of Proposition 2

Proof. We prove that the restricted mapping $\phi_q|_D : D \to \mathcal{U}_D$ is bijective by showing injectivity and surjectivity.

Injectivity. We need to show that for any $i_k, i_l \in D$, if $\phi_q(i_k) = \phi_q(i_l)$, then $i_k = i_l$. By Proposition 1, for any two distinct pages we have

$$\forall\, i_k \neq i_l, \quad \phi_q(i_k) \neq \phi_q(i_l) \tag{50}$$

Taking the contrapositive yields

$$\phi_q(i_k) = \phi_q(i_l) \implies i_k = i_l \tag{51}$$

and thus $\phi_q|_D$ is injective.

Surjectivity. We need to show that for any $u \in \mathcal{U}_D$, there exists an $i \in D$ such that $\phi_q(i) = u$. By definition,

$$\mathcal{U}_D = \{u_j\}_{j=1}^N = \{\phi_q(i_j)\}_{j=1}^N \tag{52}$$

Hence, for any $u \in \mathcal{U}_D$, there exists some $j \in \{1, \ldots, N\}$ such that

$$u = u_j = \phi_q(i_j) \tag{53}$$

Let $i = i_j \in D$. Then $\phi_q(i) = u$, proving surjectivity.

Since $\phi_q|_D$ is both injective and surjective, it is bijective. □

A.2.3 Proof of Proposition 3

Proof. We aim to show that the answer semantics u_a are contained in the evidence page's unique semantics u_{i^*}, i.e., $I(u_a; u_{i^*}) = H(u_a)$.

Since the answer a appears on page i^*, it can be viewed as information extracted from i^* under the guidance of the question q. Formally, we model the answer as a question-conditioned projection of the page content:

$$a = \pi_q(i^*) \tag{54}$$

where π_q is an information extraction operator guided by q.

Let $u_a = f^t_{\mathrm{enc}}(a, q) = f^t_{\mathrm{enc}}(\pi_q(i^*), q)$ denote the semantic representation of the answer, and let $u_{i^*} = \phi_q(i^*) = \mathrm{Encode}(i^*, q)$ denote the question-conditioned unique semantics of the evidence page. By definition,

$$I(u_a; u_{i^*}) = H(u_a) + H(u_{i^*}) - H(u_a, u_{i^*}) \tag{55}$$

Because $a = \pi_q(i^*)$ is a deterministic function of (i^*, q), we have

$$H(a \mid i^*, q) = 0 \tag{56}$$

Moreover, under the assumption that the page semantics u_{i^*} are information-preserving with respect to the answer-relevant signals (i.e., u_{i^*} deterministically determines u_a), it follows that

$$H(u_a \mid u_{i^*}) = 0 \tag{57}$$

Using the chain rule of entropy, we obtain

$$\begin{aligned} H(u_a, u_{i^*}) &= H(u_{i^*}) + H(u_a \mid u_{i^*}) \\ &= H(u_{i^*}) \end{aligned} \tag{58}$$

Substituting into the definition of mutual information yields

$$\begin{aligned} I(u_a; u_{i^*}) &= H(u_a) + H(u_{i^*}) - H(u_a, u_{i^*}) \\ &= H(u_a) + H(u_{i^*}) - H(u_{i^*}) \\ &= H(u_a) \end{aligned} \tag{59}$$

Therefore, by the definition of the containment relation $\sqsubseteq$, we conclude that

$$u_a \sqsubseteq u_{i^*} \tag{60}$$

This completes the proof. □

A.3 Additional Experimental Details

A.3.1 Implementation Details

All experiments are conducted on six NVIDIA GeForce RTX 4090 GPUs using PyTorch 2.7.1. The detailed training configurations are as follows.

Optimization Settings. We employ the Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The initial learning rate is set to 1.0×10^{-4}, and the weight decay is 1.0×10^{-5}. A cosine annealing schedule is used to dynamically adjust the learning rate, which decays according to:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos\left(\frac{t}{T}\pi\right)\right) \tag{61}$$

where $\eta_{\max} = 1.0 \times 10^{-4}$, $\eta_{\min} = 1.0 \times 10^{-6}$, t denotes the current epoch, and $T = 30$ is the total number of training epochs. Gradient clipping with a maximum norm of 1.0 is applied to stabilize training.

Model Architecture. The FUMA architecture consists of the following components:

- **Image Encoder:** We adopt the pre-trained visual model from ColQwen3 (Huang and Tan, 2025) to encode each document page, producing representations of dimension $d_{\text{model}} = 4096$.
- **Text Encoder:** We adopt the first 8 Transformer blocks from the language model of ColQwen3 (Huang and Tan, 2025) to encode questions and answers, producing representations of dimension $d_{\text{model}} = 4096$.
- **Flow-Based Blocks:** The number of stacked Flow-Based Blocks is set to $m = 6$. Each block contains multi-head attention with $H = 8$ heads and feed-forward networks with intermediate dimension $d_{ff} = 1024$.
- **Multi-Expert Decoder:** The total number of experts is set to $N = 4$, with each expert having the same architecture as a single Flow-Based Block. During inference, we select the top-$K = 2$ experts based on gating scores.

Training Strategy. The batch size is set to 16 with gradient accumulation over 2 steps, resulting in an effective batch size of 32. The model is trained for up to 30 epochs, and early stopping is applied based on validation performance: training is halted if the validation loss does not improve for 5 consecutive epochs. The bijective regularization weight in the joint loss function is set to $\gamma = 0.1$.

During training, we adopt mixed-precision computation (BF16) to reduce memory usage and accelerate training.

Hyperparameter Selection. Key hyperparameters were selected through grid search on the validation set. We conducted detailed ablation studies on the NiM-Benchmark validation set. The results demonstrate that inappropriate hyperparameter choices can significantly degrade model performance, highlighting the critical importance of hyperparameter tuning.

Selection of Flow-Based Block Number. Table 7 shows the impact of different numbers of Flow-Based Blocks on model performance. When $m = 4$, the model achieves 79.24% EM, 88.53% F1, and 91.67% ANLS, indicating that insufficient network depth limits the model's capacity to learn complex bijective mappings between pages

and their unique semantics; at $m = 6$, performance reaches the optimal 83.57% EM, 92.79% F1, and 95.44% ANLS, demonstrating that moderate depth effectively captures fine-grained semantic distinctions while maintaining stable gradient flow; however, at $m = 8$, performance decreases to 82.18% EM, 91.05% F1, and 93.82% ANLS (drops of 1.39, 1.74, and 1.62 percentage points respectively), while inference time increases by 35% (from 189ms to 255ms). This indicates that excessively deep flow-based architectures not only substantially increase computational cost but also suffer from overfitting and gradient propagation issues, which degrade the quality of unique semantic representations. Therefore, we select $m = 6$ as the optimal balance between representation capacity and computational efficiency.

Selection of Expert Number and Top-K. Table 8 shows the impact of different combinations of total expert number N and Top-K selection on model performance. For each expert number $N \in \{4, 6, 8\}$, we systematically tested all possible values from $K = 1$ to $K = N$. The results reveal clear patterns: (1) When $N = 4$, EM is only 78.42% for $K = 1$, reaches the optimal 83.57% at $K = 2$ (improvement of 5.15 percentage points), but declines as K continues to increase, dropping to 81.29% for $K = 3$ and further to 79.63% for $K = 4$ (using all experts); (2) When $N = 6$, EM is 76.85% for $K = 1$, improves to 81.94% for $K = 2$, reaches the peak of 82.47% at $K = 3$, but decreases to 81.16% for $K = 4$ and continues to decline for $K \geq 5$, dropping to 79.15% at $K = 6$; (3) When $N = 8$, performance gradually improves from $K = 1$ to $K = 3$ (from 75.32% to 81.83% EM), but continuously declines for $K \geq 4$ (dropping to 78.01% EM at $K = 8$), while inference time significantly increases (from 158ms to 309ms). These results indicate that: (i) using a single expert ($K = 1$) cannot fully exploit multimodal complementarity, resulting in 5 percentage point performance drops across different N values; (ii) more experts are not always better—while larger expert pools ($N = 6, 8$) provide greater capacity, they also introduce training instability and redundant representations; (iii) selecting too many experts (K close to or equal to N) introduces noisy information and dilutes the contribution of truly relevant experts, degrading answer quality. Overall, the combination ($N = 4, K = 2$) achieves the best balance between performance (83.57% EM, 92.79% F1, 95.44% ANLS) and efficiency (155ms inference time), ensuring expert diversity and complementarity while avoiding the redundancy and computational overhead of excessive experts.

Selection of Bijective Regularization Weight. Table 9 shows the impact of different bijective regularization weights on model performance. When $\gamma = 0.01$, the regularization effect is too weak to enforce sufficient invertibility constraints, resulting in substantial performance degradation: EM drops to 68.34%, F1 to 76.82%, and ANLS to 80.15% (losses of 15.23, 15.97, and 15.29 percentage points respectively compared to the optimum). At this setting, the model tends to collapse into traditional similarity-based matching, failing to establish well-separated unique semantic representations for different pages. At $\gamma = 0.1$, performance reaches the optimal 83.57% EM, 92.79% F1, and 95.44% ANLS, effectively balancing bijective invertibility constraints with task-oriented discriminative learning. At $\gamma = 0.5$, performance decreases to 71.48% EM, 80.26% F1, and 83.67% ANLS (losses of 12.09, 12.53, and 11.77 percentage points respectively), indicating that over-regularization begins to constrain the model's expressiveness. At $\gamma = 1.0$, performance further plummets to 63.91% EM, 72.15% F1, and 75.89% ANLS (losses of 19.66, 20.64, and 19.55 percentage points respectively), demonstrating that excessive regularization severely restricts the model's capacity to learn discriminative unique semantics, as the training process becomes dominated by invertibility constraints rather than evidence localization objectives. The clear inverted-U pattern across all metrics confirms that $\gamma = 0.1$ represents the optimal operating point, where bijective regularization effectively prevents representation collapse while preserving sufficient flexibility for task-specific semantic learning.

A.3.2 Dataset Details

We evaluate FUMA on four benchmark datasets covering diverse document types and varying levels of complexity.

(1) NiM-Benchmark This dataset consists of 2,970 document page images and 1,180 question–answer pairs, spanning various document types including academic papers, newspapers, menus, and lecture slides. Each document contains an average of 2.5 pages, where questions require understanding both textual content and visual elements such as tables and figures.

Table 7: Impact of the number of Flow-Based Blocks on model performance. Results on the NiM-Benchmark validation set.

m	**EM (%)**	**F1 (%)**	**ANLS (%)**	**Inference Time (ms)**
4	79.24	88.53	91.67	145
6	**83.57**	**92.79**	**95.44**	189
8	82.18	91.05	93.82	255

Table 8: Impact of different combinations of total expert number N and Top-K selection on model performance. Results on the NiM-Benchmark validation set.

N	K	**EM (%)**	**F1 (%)**	**ANLS (%)**	**Inference Time (ms)**
4	1	78.42	87.16	90.28	142
4	2	**83.57**	**92.79**	**95.44**	155
4	3	81.29	90.45	93.17	172
4	4	79.63	88.82	91.56	185
6	1	76.85	85.73	88.94	152
6	2	81.94	90.68	93.51	165
6	3	82.47	91.23	94.05	189
6	4	81.16	89.91	92.68	218
6	5	80.28	89.04	91.82	241
6	6	79.15	87.95	90.73	258
8	1	75.32	84.28	87.56	158
8	2	80.71	89.45	92.18	178
8	3	81.83	90.52	93.34	230
8	4	81.45	90.12	92.89	250
8	5	80.67	89.38	92.15	262
8	6	79.82	88.54	91.28	275
8	7	78.94	87.69	90.42	291
8	8	78.01	86.82	89.54	309

Table 9: Impact of bijective regularization weight on model performance. Results on the NiM-Benchmark validation set.

γ	**EM (%)**	**F1 (%)**	**ANLS (%)**
0.01	68.34	76.82	80.15
0.1	**83.57**	**92.79**	**95.44**
0.5	71.48	80.26	83.67
1.0	63.91	72.15	75.89

(2) DUDE The Document Understanding Dataset (DUDE) comprises 41491 question–answer pairs from 5,019 documents across five domains: scientific articles, financial reports, government tenders, medical forms, and legal contracts. Each document has an average of 5.72 pages. Due to its domain diversity and the need for cross-page information retrieval, DUDE poses significant challenges.

(3) MMLongBench-Doc This benchmark focuses on long documents with an average length of 47.5 pages. It includes 135 documents and 1,082 question–answer pairs, designed for documents containing extensive text, complex charts, graphs, and schematic illustrations. The documents are collected from academic papers, technical reports, and manuals. MMLongBench-Doc challenges models to handle extremely long contexts and maintain semantic consistency across distant pages.

(4) DocVQA The Document Visual Question Answering (DocVQA) dataset contains 12000+ document images and 50,000 questions, sourced from the UCSF Industry Document Library. It includes real-world documents such as letters, forms, invoices, advertisements, and reports. Unlike multi-page datasets, DocVQA focuses on single-page understanding but requires fine-grained visual reasoning.

A.3.3 Baseline Methods

We compare FUMA with the following state-of-the-art models:

- **Hi-VT5**:Hi-VT5 is a multimodal hierarchical Transformer model that summarizes key information from each page in a bottom-up manner via the encoder and generates the final answer through the decoder, capable of simultaneously performing multi-page document question answering and answer page localization in a single stage.
- **LayoutLMv3**: LayoutLMv3 is a multimodal Transformer pre-trained model for Document AI that learns cross-modal representations through unified text masking (MLM), image masking (MIM), and Word-Patch Alignment (WPA) pre-training objectives.
- **DocFormer**: DocFormer is a multimodal Transformer document understanding model that fuses text, vision, and spatial features through a multi-modal self-attention layer, and shares spatial embeddings across modalities to correlate text and visual tokens.
- **GRAM**: GRAM is a multi-page document question answering model that builds upon a pre-trained single-page encoder by adding document-level designated layers and learnable tokens to enable cross-page information flow, and employs a bias adaptation method to guide the model to utilize document tokens.
- **MP-FIRE**: MP-FIRE is a multi-page document question answering framework that extracts document features using a topology graph and applies a two-stage pruning process to eliminate irrelevant elements, combined with a cross-modal agent ensemble based on the Performance Characterization Spectrum (PCS).

For fair comparison, all baseline methods are trained under the same optimization settings on each dataset.

A.3.4 Evaluation Metrics

We adopt three standard metrics to comprehensively evaluate model performance.

Exact Match (EM). EM measures the percentage of predictions that exactly match the ground-truth answers, defined as:

$$\text{EM} = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \mathbb{I}[\text{pred}_i = \text{gt}_i] \tag{62}$$

where $\mathcal{Q}$ is the set of questions, pred_i is the predicted answer, gt_i is the ground truth, and $\mathbb{I}[\cdot]$ denotes the indicator function. EM is a strict metric that rewards only exact matches.

F1 Score. F1 computes the harmonic mean of token-level precision and recall:

$$\text{Precision} = \frac{|\text{pred} \cap \text{gt}|}{|\text{pred}|} \tag{63}$$

$$\text{Recall} = \frac{|\text{pred} \cap \text{gt}|}{|\text{gt}|} \tag{64}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{65}$$

where $|\cdot|$ denotes the number of tokens. F1 provides a softer evaluation that rewards partial matches between predictions and ground truths.

Average Normalized Levenshtein Similarity (ANLS). ANLS evaluates similarity based on edit distance and is more robust to minor variations:

$$\text{NLS}(s_1, s_2) = 1 - \frac{\text{Lev}(s_1, s_2)}{\max(|s_1|, |s_2|)} \tag{66}$$

$$\text{ANLS} = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \max_j \text{NLS}(\text{pred}_i, \text{gt}_{i,j}) \tag{67}$$

where $\text{Lev}(\cdot, \cdot)$ denotes the Levenshtein distance. When multiple reference answers are available, the maximum similarity is taken. ANLS ranges from 0 to 1, with higher values indicating better performance.