DocGenie: A Framework for High-Fidelity Synthetic Document Generation via Seed-Guided Multimodal LLM and Document-Aware Evaluation

Anonymous CVPR submission

Paper ID *****

Abstract

Obtaining large-scale, high-quality datasets for document understanding tasks such as optical character recognition, key information extraction, and layout analysis is costly and time-consuming. Synthetic document generation offers a scalable alternative, but achieving visual realism, structural coherence, and semantic alignment remains a challenge. This work presents DocGenie, a framework for generating high-fidelity, domain-adaptable synthetic business documents using a frontier multimodal large language model (MLLM). DocGenie leverages seed examples to guide HTML-based document generation, aligning outputs with domain-specific content and layout conventions. To evaluate quality and similarity, DocGenie introduces Layout-FID, a document-aware adaptation of Fréchet Inception Distance that replaces InceptionV3 with LayoutLMv3 embeddings. Layout-FID better captures textual, structural, and visual features, yielding more reliable scores across various business document categories: invoices, receipts, forms, and budgets. To enhance the visual realism of the generated documents, two post-processing strategies are explored: distortions derived from seed documents via (i) human inspection and (ii) MLLM-based prediction. This comparative study assesses their effectiveness across document categories with varying realistic distortion profiles. DocGenie thus offers a practical and extensible solution for realistic synthetic document generation and evaluation tailored for document AI workflows.

1. Introduction

Document understanding plays a central role in various AI applications, including optical character recognition (OCR), key information extraction, layout analysis, and classification. Visual document understanding (VDU) models [17, 19, 20, 30] have demonstrated that extensive pretraining with document corpora is essential for state-of-the-art performance. However, acquiring and annotating real-world documents is costly, labor intensive and constrained by privacy concerns [8].

Synthetic document generation has emerged as a scalable alternative to real-world data collection. However, generating realistic and diverse documents remains a challenging task. Existing approaches such as rule-based templates [31], GAN-based synthesis [3], diffusion model-based generation [28], and LLM-driven generation [4] - have been largely developed and evaluated on datasets like PubLayNet [31], which primarily consist of scientific articles with structured layouts. Also, they often lack visual realism, structural coherence, or adaptability to specific document domains.

To address the limitations of existing synthetic document generation methods, this work introduces **DocGenie**, a framework for generating high-fidelity, domain-adaptable synthetic business documents using a multimodal large language model (MLLM). Here, high-fidelity refers to the generation of documents that are visually realistic, structurally coherent, and semantically aligned with the target domain [23, 24]. These properties are critical for training document understanding models that rely on synthetic data to generalize effectively to real-world scenarios. DocGenie supports seed-guided generation to enhance domain alignment and structural consistency. It targets the synthesis of business documents such as invoices, receipts, forms, and budgets, which are typically semi-structured and require semantically rich, domain-specific content along with coherent layout and visual realism. The framework also adapts the Fréchet Inception Distance (FID) [9] for document evaluation by replacing InceptionV3 with LayoutLMv3 embeddings [10], resulting in Layout-FID, a metric better suited to capture the structural, textual, and visual properties of documents.

In addition, to investigate how document visual realism could be effectively incorporated within the DocGenie framework, a comparative study is conducted in which two document distortion selection strategies are explored: (i) distortions manually identified through human inspection of real seed documents, and (ii) distortions dynami-

cally predicted by an MLLM based on the same seed inputs. The study reveals that while the MLLM-driven strategy performs reasonably well when seed documents contain significant noise, it tends to hallucinate distortions when seed samples are clean. In contrast, human-curated distortions yield more reliable and consistent improvements across document categories.

To the best of our knowledge, this is the first work to propose a unified framework for generating high-fidelity synthetic business documents across diverse formats. Although prior work has addressed synthetic generation in structured domains such as scientific articles or webpages to some extent, no existing baselines target the visual, structural, and semantic variability found in business documents.

The key contributions of DocGenie are as follows:

- A scalable framework for MLLM-driven generation of structured HTML-based documents, capable of capturing both content and layout. The framework supports seed-guided generation, which enhances domain adaptability by aligning outputs with specific formats, styles, and vocabulary.
- A document-specific evaluation strategy based on a modified FID, referred to as Layout-FID, which replaces the InceptionV3 backbone with LayoutLMv3 [10] embeddings. This metric jointly captures textual, visual, and spatial properties, offering a more meaningful measure of document similarity than conventional image-based FID.
- A comparative study of visual realism enhancement strategies applied on generated documents, evaluating both human-curated distortions and distortions dynamically predicted by an MLLM. The analysis reveals limitations in the MLLM-based approach when seed documents lack significant noise, highlighting the need for further research into adaptive and content-aware realism mechanisms for synthetic document generation.

2. Related Works

Synthetic document generation has emerged as an efficient alternative to labor-intensive real-world data collection. Traditional approaches rely on rule-based template generation, where textual elements are arranged in predefined layouts [6, 17, 26, 31]. While these methods offer structural consistency and control, they often lack semantic diversity and visual variability, making them less suitable for modeling the wide range of layouts and content found in real-world business documents.

To improve diversity and realism, deep generative models such as GANs [3, 7] and diffusion-based methods [12, 29] have been explored. These approaches typically focus on layout generation, often requiring additional modules for text infilling and complex training pipelines. Furthermore, most of them are evaluated on datasets like PubLayNet [31], which predominantly contain scientific or academic documents with rigid structured layouts limiting their applicability to semi-structured business documents such as invoices, receipts, and forms.

Recent LLM-powered approaches treat document layout generation as a structured code generation task. Works like LAYOUTNUWA [27] and PosterLlama [25] leverage instruction tuning to produce layout code but fall short of full document synthesis. DocSynth2 [4] moves closer to end-to-end generation by modeling both layout and content jointly through a GPT-2-based decoder. However, it is primarily evaluated on structured academic layouts and lacks domain adaptability. In contrast, DocGenie targets business document generation using a frontier multimodal LLM, enabling end-to-end HTML-based content and layout generation without fine-tuning. Furthermore, seed samples can be optionally provided to guide domain-specific structure and language, improving alignment and diversity across document types.

Evaluating synthetic document quality remains a fundamental challenge. Fréchet Inception Distance (FID) [9] is widely used but is designed for natural images and fails to capture the multimodal characteristics of documents [15]. Metrics like mIoU and overlap scores [13] evaluate layout alignment but do not account for text or visual realism. OCR-based accuracy measures [16] provide indirect signals but lack holistic assessment. To address this, DocGenie introduces Layout-FID, an FID adaptation that replaces the InceptionV3 embedding model with LayoutLMv3 [10], enabling a more robust similarity comparison that jointly reflects layout structure, text content, and visual features. This document-specific adaptation bridges a critical gap in synthetic document evaluation.

A key challenge in synthetic document generation is ensuring that generated samples not only preserve structural and semantic fidelity but also reflect the visual realism of real-world documents. Realism is particularly important for downstream tasks such as OCR and information extraction, where model performance can be sensitive to surface-level artifacts. To this end, several prior works have applied image-based techniques to simulate real-world imperfections. Methods such as [16, 17, 26] use hand-crafted visual effects that include font deformation, background blending, and scanning noise to increase authenticity. However, these effects are typically applied statically, without regard for whether such imperfections are present in the target domain or seed samples, often resulting in realism cues that are misaligned with the expected document style. GAN-based methods [5] attempt to learn visual degradations such as blur, compression, and lighting variation, but require additional training and lack controllability. Despite growing interest, few existing works have systematically studied how visual realism should be introduced in synthetic document generation pipelines. To address this, Doc-

Genie conducts a comparative study of two strategies for enhancing visual realism in synthetic documents, informed by characteristics observed in seed examples: (i) manually identified realism cues derived through domain-specific inspection of real documents, and (ii) realistic effects dynamically suggested by a MLLM conditioned on the seed inputs. This analysis underscores the importance of a document-aware, content-driven realism layer moving beyond static augmentation recipes toward adaptive techniques that reflect domain-specific visual characteristics.

3. DocGenie Framework: Seed-Guided Generation and Document-Aware Evaluation

3.1. Overview of DocGenie

DocGenie is a modular framework for high-fidelity synthetic document generation, designed to produce visually realistic, structurally coherent, and semantically aligned business documents across multiple domains. It consists of three core components: **(1)** seed-guided HTML-based document generation using a frontier multimodal large language model (MLLM), **(2)** document-specific quality evaluation using a novel metric called Layout-FID, and **(3)** an optional visual realism enhancement module to explore strategies to further improve visual realistic aspects.

The key novelty of DocGenie lies in its **seed-guided generation mechanism**, where a small set of real documents is used to guide the MLLM in producing domain-aligned, culturally appropriate, and layout-consistent synthetic samples. Unlike template-based or layout-only methods, DocGenie generates semantically rich and structurally diverse HTML documents in a single step. Another major contribution is the introduction of **Layout-FID**, a document-aware adaptation of the Fréchet Inception Distance (FID), to better capture textual, visual, and spatial structure making it more suitable for document generation evaluation tasks.

Additionally, we explore the role of post-hoc **visual realism enhancement** through a comparative study of two distortion selection strategies: (i) human-guided and (ii) MLLM-guided. This component is not part of the core generation-evaluation loop, but serves to investigate how realism artifacts can be introduced to bridge the synthetic-real gap in visually noisy domains. Regardless of whether realism is applied, the Layout-FID metric consistently evaluates the alignment between the generated and seed documents.

An overview of the DocGenie framework is shown in Figure 1.

3.2. Synthetic Document Generation with MLLM

At the core of DocGenie is a frontier MLLM that generates documents in structured HTML code format. The model is conditioned on a user-specified document category (e.g., receipts, invoices, forms) and a set of seed samples of the same category, which serve as reference points for aligning the style, structure, content type, and domain-specific patterns in the output. Seed samples are a critical component of the generation process, guiding the MLLM to produce semantically and culturally aligned outputs that reflect real-world document conventions.

Let:

- $S = \{s_1, s_2, ..., s_m\}$ be the set of seed document samples, each s_i representing a real document instance.
- C be the target document category (e.g., invoice, budget).
- N be the desired number of synthetic documents.
- Φ be the multimodal LLM that maps the seed set and category to structured outputs.

The generation process yields a set of synthetic HTML-based documents $\hat{D} = \{\hat{d}_1, \hat{d}_2, ..., \hat{d}_N\}$, defined as:

$$\hat{D} = \Phi(S, C, N) \tag{1}$$

The underlying conditional probability is modeled as:

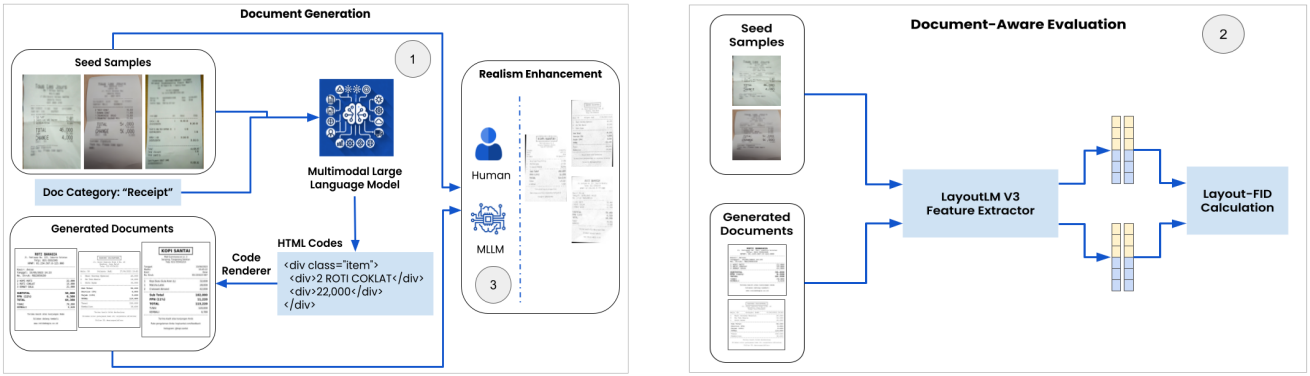
$$P(\hat{D}|S, C) = \prod_{i=1}^{N} P(\hat{d}_i|S, C) \tag{2}$$

This setup enables the MLLM to generate documents that are diverse in layout and content but anchored in the structure and semantics of the seed samples. The HTML outputs are subsequently rendered as PDFs to simulate realistic document appearances.

3.3. Layout-FID: A Document-Centric Adaptation of FID

To evaluate the quality, similarity and alignment of the generated documents with the seed samples, DocGenie uses Layout-FID, a modification of the Fréchet Inception Distance (FID) that replaces the InceptionV3 backbone with LayoutLMv3 embeddings [10]. FID has been widely adopted for evaluating generative models due to its ability to quantify both fidelity and diversity on unlabeled data, its robustness to visual corruption, and its strong correlation with perceptual quality [24, 28]. A lower value indicates that the generated images are more similar to the real images. However, its reliance on natural image embeddings makes it unsuitable for document specific evaluation.

LayoutLMv3, by contrast, is pre-trained on large-scale document corpora and produces unified embeddings that jointly capture textual content, visual features, and spatial layout. Its effectiveness has led to its adoption in recent document understanding frameworks, such as LayoutLLM [21], where it serves as the encoder backbone. Leveraging these properties, Layout-FID offers a more appropriate and domain aligned metric for assessing the structural, semantic, and stylistic fidelity of synthetic documents. Layout-FID is computed between the LayoutLMv3 embeddings of



(a) (b)

Figure 1. Overview of the DocGenie Framework: (a) Seed-guided MLLM-based generation and optional realism enhancement, (b) LayoutLMv3 based evaluation comparing seed and generated documents.

the generated documents and their corresponding seed samples, enabling a document-specific evaluation grounded in domain-relevant characteristics.

Formally, the Layout-FID is calculated as:

$$\text{Layout-FID} = ||\mu_r - \mu_g||^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}) \quad (3)$$

where:

- μ_r, Σ_r: mean and covariance of LayoutLMv3 embeddings from the seed documents.
- μ_g, Σ_g: mean and covariance of LayoutLMv3 embeddings from the generated documents.

This evaluation strategy offers a robust, multimodal comparison mechanism, distinguishing it from conventional FID, mIOU, or OCR-only metrics that fail to capture the full structural and semantic richness of documents.

3.4. Evaluating Visual Realism Strategies for DocGenie

In addition to its core generation and evaluation modules, this work presents a comparative study on enhancing the visual realism of synthetic documents, examining post-processing strategies that can optionally complement the DocGenie framework. While the HTML-rendered outputs are structurally coherent, they may lack the natural imperfections typically present in real scanned or photographed documents. To determine effective strategies for incorporating realism, a comparative analysis is conducted using a predefined set of common visual artifacts such as blur, compression, background noise, shadowing, etc as outlined in the Donut paper [17]. This list serves as a reference taxonomy of visual effects frequently found in real documents. Following are the list of visual distortions considered.

- **Background Blending** – Adds blurred textures to eliminate artificial white space.
- **Paper Texturing** – Overlays natural paper textures for realistic surfaces.
- **Visual Distortions & Noise** – Applies Gaussian noise, Salt and Pepper noise, elastic deformations, and perspective shifts.
- **Compression Artifacts & Blurring** – Simulates JPEG compression, motion blur, and scan imperfections.
- **Shadow & Vignetting** – Introduces lighting variations through radial, linear, or irregular shadows.

Given this reference set, two strategies are used to select which realism effects to apply to the synthetic outputs:

1. **Human-Curated Realism:** Human annotators inspect the seed samples to identify which of the listed artifacts are present in the given document category. Only the matched effects are then applied to the generated documents.
2. **MLLM-Predicted Realism:** A multimodal LLM is prompted with the same seed documents and the artifact list, and asked to identify which visual effects are appropriate to add based on the observed characteristics of the seeds.

The selected visual effects are applied to the rendered synthetic documents as a post-processing step. To evaluate the impact of realism addition, Layout-FID is computed between the seed documents and both the original and realism enhanced synthetic documents.

4. Experiments

4.1. Experimental Setup

To assess the effectiveness and generalizability of DocGenie, experiments are conducted across five structurally and visually diverse business document categories: CORD receipts [22], SROIE receipts [11], budgets from the RVL-CDIP dataset [8], forms from the Doc-UFC competition

[14], and electronic invoices from a publicly available Mendeley repository [18]. These categories capture a spectrum of real-world document complexity, layout variation, and visual degradation. The visual quality varies across types, with degradation approximately increasing in the following order: Invoice → Form → Budget → SROIE Receipt → CORD Receipt.

Document Generation with Claude-Sonnet 3.7. Claude-Sonnet 3.7 [2] is employed as the underlying MLLM for HTML-based document generation. For each document category, a set of $S = 10$ real documents is selected as seed samples to guide the generation process. The MLLM is prompted with the seed samples and document category to generate $N = 10$ synthetic documents per category. Each model call generates 4 HTML-based documents per iteration, repeated until the total target is reached. All generated HTML documents are rendered into high-resolution PDF format using a rendering engine [1]. This ensures the preservation of layout, font styling, and spacing necessary for reliable evaluation.

Seed-Free Baseline. To evaluate the influence of seed guidance, synthetic documents are also generated without providing seed samples (i.e., zero-shot generation). This enables a direct comparison of domain alignment and structure between seed-guided and seed-free generations.

Seed-vs-Seed Baseline. To establish an upper-bound reference score for similarity within the same domain, FID is also computed between two disjoint subsets of the seed documents themselves. This provides a sanity check for intra-domain variation and helps contextualize how closely synthetic documents approach the distribution of real ones.

Evaluation Protocol. We compute both **Inception-FID** and **Layout-FID** to compare traditional image-based evaluation with our proposed document-specific metric. Inception-FID uses InceptionV3 embeddings and serves as a baseline, while Layout-FID, calculated using LayoutLMv3 embeddings [10] captures textual, visual and structural alignment relevant to document tasks. This setup allows us to assess the effectiveness of Layout-FID in measuring the quality of synthetic documents.

Evaluation Stages. All FID scores are reported in two settings:

1. *Raw Output:* Synthetic documents directly rendered from the MLLM's HTML outputs.
2. *Realistic Output:* Synthetic documents post-processed using visual realism strategies (described in Section 4.3).

4.2. Seed-Guided vs. Seed-Free Generation

To evaluate the impact of seed guidance on the quality of synthetic document generation, we compare outputs generated with and without seed samples across all five document categories. The goal is to measure how reference samples affect structural alignment, semantic fidelity, and domain specificity in generated documents.

Table 1. Inception-FID and Layout-FID scores across five document categories for seed-free and seed-guided generation, along with the baseline computed between real seed samples (Seed vs. Seed). Lower scores indicate higher similarity.

Setting	Inception-FID ↓	Layout-FID ↓
CORD		
Seed vs. Seed	160.062	6.120
Seed-Free Generation	207.615	53.242
Seed-Guided Generation	155.335	31.299
SROIE		
Seed vs. Seed	70.284	3.483
Seed-Free Generation	147.463	7.059
Seed-Guided Generation	109.307	3.524
Invoice		
Seed vs. Seed	8.509	0.875
Seed-Free Generation	51.444	9.171
Seed-Guided Generation	40.842	5.614
Form604		
Seed vs. Seed	26.029	1.460
Seed-Free Generation	84.616	10.881
Seed-Guided Generation	44.748	3.448
Budget		
Seed vs. Seed	90.948	11.125
Seed-Free Generation	137.279	15.289
Seed-Guided Generation	87.152	8.757

For each document type, a fixed set of $S = 10$ seed samples is selected. Using Claude-Sonnet 3.7, we perform three independent generation runs to produce $N = 10$ synthetic documents per category in both seed-guided and seed-free modes. This setup accounts for the stochastic nature of LLM output while keeping the seed set constant. All generated documents are rendered into PDFs and evaluated using both Inception-FID and Layout-FID.

We also compute a baseline FID score by comparing two non-overlapping subsets of the seed samples (seed-vs-seed). This score reflects the inherent intra-domain variation and acts as an upper-bound reference to assess how closely the synthetic outputs resemble real documents from the same distribution.

As shown in Table 1, seed-guided generation consistently yields lower Layout-FID scores across all document categories, indicating stronger alignment with the structural and semantic characteristics of the seed domain. In document types like Form604 and Invoice (Figure 2 column 3,4), where the layout is highly consistent and only textual content varies, the seed-guided Layout-FID scores (3.448 and 5.614) approach the seed-vs-seed baselines (1.460 and 0.875), reflecting strong alignment. In contrast, domains with higher intra-category variation—such as Budgets and

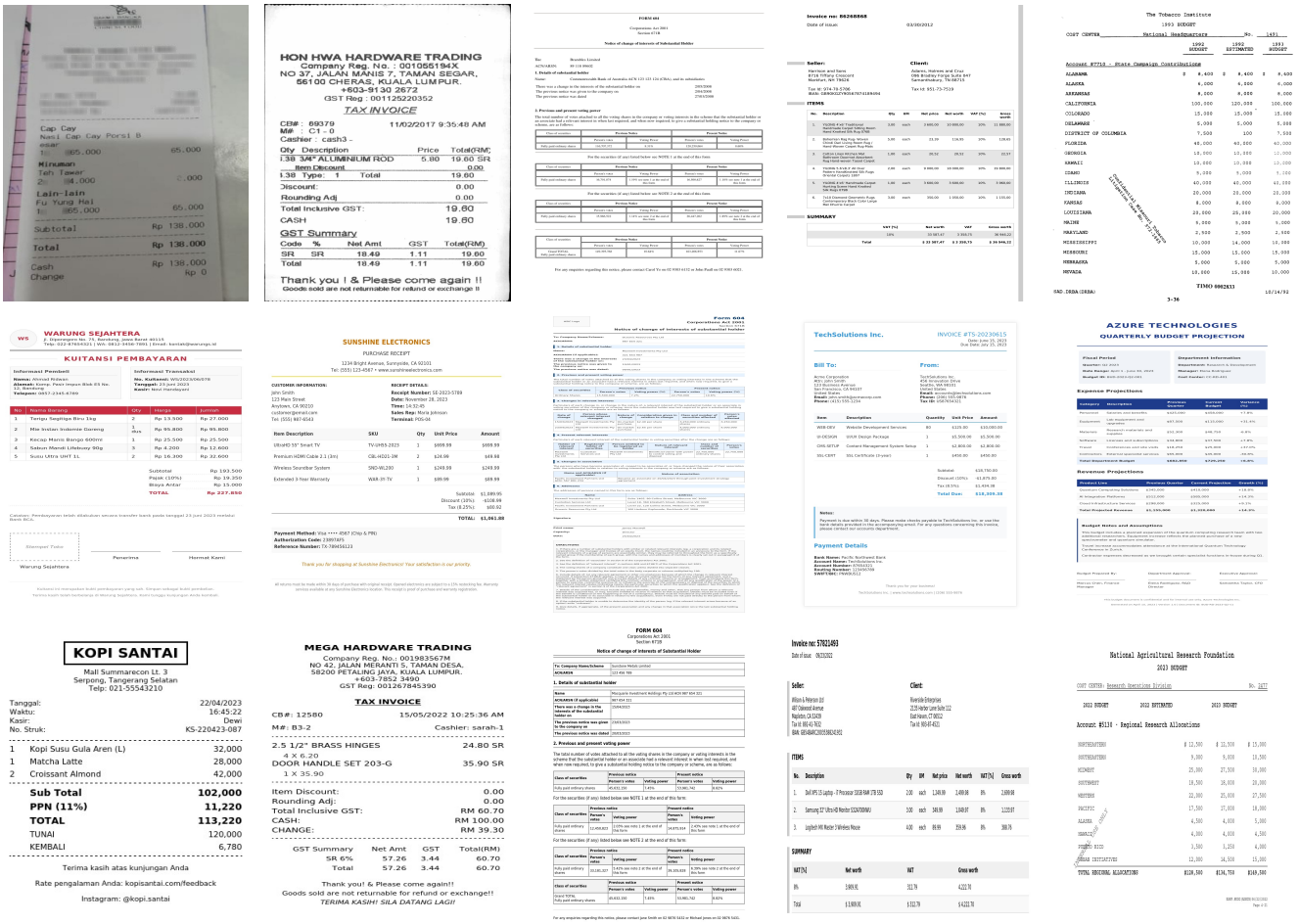


Figure 2. Qualitative comparison across document types. Columns correspond to document categories: CORD, SROIE, Form604, Invoice, and Budget. Each row shows a real seed sample (top), a seed-free synthetic generation (middle), and a seed-guided generation (bottom). All document types show notable improvements in structure and fidelity with seed guidance.

Receipts (CORD) show larger absolute Layout-FID values, but still exhibit a substantial improvement when seeds are used. The qualitative examples in Figure 2 support these observations, showing that seed-guided outputs maintain domain-specific layout and semantic conventions more faithfully.

These results also underscore the limitations of Inception-FID for document-level evaluation. While the relative trend of score reduction across settings (e.g., seed-free vs. seed-guided) appears directionally consistent with Layout-FID, the absolute values of Inception-FID remain disproportionately high even for visually clean and structurally consistent categories like Invoice and Form604. This suggests that Inception embeddings fail to capture fine-grained textual, layout, and structural similarities that are critical to document understanding. In contrast, Layout-FID built on LayoutLMv3, a document-pretrained encoder offers a more reliable and interpretable metric. Its lower absolute scores reflect a higher degree of semantic, structural, and visual alignment with the seed samples, supporting its suitability for evaluating synthetic document fidelity beyond what natural image-based metrics can offer. The rise in Layout-FID scores under seed-free generation clearly indicates a lack of alignment with cultural, linguistic, and domain-specific conventions. Without access to representative examples, the MLLM generates structurally reasonable but semantically generic documents that deviate from the stylistic norms of each domain. However, the performance gap observed typically within 5-15 Layout-FID points also shows that DocGenie retains a strong intrinsic prior, enabling it to generate coherent outputs even in the absence of explicit seed samples. This highlights the framework's applicability in zero-shot settings where real seed data are unavailable while reinforcing the added value of seed guidance for precise domain adaptation.

4.3. Visual Realism Enhancement Study

This experiment evaluates the impact of post-processing strategies for adding visual realism to synthetic documents. The goal is to assess whether guided realism based on distortions observed in real seed samples improves alignment with real documents as measured by Layout-FID.

We compare two strategies: (i) **Human-Guided Realism Enhancement**, where distortions are manually identified based on visual inspection of the seed documents, and (ii) **MLLM-Guided Realism Enhancement**, where distortions are predicted by Claude-Sonnet 3.7 based on the same visual inputs.

For each document category, we select a set of 10 seed samples. We reference the list of visual artifacts from the Donut paper, including background blending, compression artifacts, Gaussian noise, motion blur, shadowing, and perspective shifts. Each distortion type is assessed independently.

Three independent human experts were tasked with grading the presence of each distortion as *high*, *medium*, or *low*. The final severity level per distortion type was decided via majority vote. The same process was replicated using Claude-Sonnet 3.7, which received the seed documents and the predefined distortion list and returned its predicted severity levels.

Based on the chosen severity levels, OpenCV based functions were used to apply each distortion to the synthetic documents. Each severity level maps to a numeric parameter range (e.g., blur kernel size, noise standard deviation, compression quality). Once a severity level is selected, the final distortion parameters are randomly sampled from that range. This preserves consistency with human or MLLM annotations while introducing realistic variability.

We then compute Layout-FID between the real seed documents and synthetic documents under three conditions: (1) raw output with no realism, (2) realism added based on human annotations, and (3) realism added based on MLLM predictions.

Table 2. Layout-FID scores between seed samples and synthetic documents across three settings: raw output, human-guided realism, and MLLM-guided realism. Lower scores indicate higher similarity.

Dataset	Raw	Human-Guided ↓	MLLM-Guided ↓
CORD	31.299	18.706	20.446
SROIE	3.524	3.157	9.411
Invoice	5.614	6.240	12.229
Form604	3.448	5.017	9.137
Budget	8.757	10.001	14.351

As shown in Table 2, the impact of visual realism enhancement varies significantly across document categories and strategies. In noisy domains such as CORD, where seed samples often contain strong real-world artifacts like blur, compression, or shadowing, both human-guided and MLLM-guided realism improve alignment with the seed distribution. Human-guided distortions reduce Layout-FID from 31.299 to 18.706, and MLLM-guided realism also lowers the score to 20.446, validating that realism artifacts can be effectively identified and applied when the seed samples exhibit such features.

However, for relatively cleaner domains such as SROIE, Budget, Form604, and Invoice the benefits of realism addition vary. Human-guided realism generally maintains or slightly improves alignment, as seen in SROIE (3.524 → 3.157), and causes only minor degradation in Form604 and Invoice, indicating its conservative and seed-aligned behavior. In contrast, MLLM-guided realism degrades performance in all these domains, significantly increasing Layout-FID due to over-application of distortions. For instance, in SROIE, Layout-FID rises to 9.411, in Invoice from 5.614 to 12.229, and in Budget, it jumps to 14.351, indicating that the MLLM tends to hallucinate non-existent artifacts when seed samples are relatively clean.

Overall, these findings reinforce that while MLLM-guided realism can be beneficial in highly degraded domains, it lacks robustness in structured and cleaner document categories. Human-guided realism, despite being labor-intensive, proves more stable across domains. These results highlight the need for improved, distortion-aware realism strategies that can dynamically adapt to the visual characteristics of the seed data.

Qualitative visual comparisons across document categories—showing outputs with no realism, human-guided realism, and MLLM-guided realism—are provided in Appendix 9.

5. Insights, Limitations, and Future Directions

The experiments presented with DocGenie highlight several key strengths and challenges in scalable synthetic document generation. First, seed-guided generation clearly enhances alignment with real-world documents across layout, linguistic structure, and cultural conventions. Layout-FID scores consistently improved with seed usage, and qualitative examples demonstrated better domain conformity in receipts, invoices, and forms. However, DocGenie also demonstrated the ability to generate reasonably aligned outputs even in seed-free scenarios, showcasing its robustness when real seed data is unavailable.

Layout-FID emerged as a more sensitive and document-aware evaluation metric than traditional Inception-FID. Its ability to capture multimodal structure textual content, spatial layout, and visual cues makes it well-suited for tasks in document understanding.

In studying visual realism strategies, human-guided en-

hancements reliably improved similarity scores across all domains. In contrast, MLLM-guided realism showed mixed results: while effective on noisy domains like CORD, it often hallucinated artifacts in clean domains, leading to worse Layout-FID.

Despite its strengths, the use of MLLMs for multi-sample HTML generation introduces practical constraints. Given the token-intensive nature of HTML and typical context length limits of large language models (e.g., 8k–16k tokens), only a limited number of documents can be generated per call in the current setup. To generate the target of ten samples per category, multiple independent calls are made without memory of previously generated content. This stateless approach occasionally results in layout repetition or content redundancy, especially in structurally rigid domains. As a future enhancement, integrating a feedback mechanism where previously generated samples are included in subsequent prompts could encourage greater diversity while preserving domain alignment. Qualitative examples of such layout repetition are shown in Appendix 10.

Additionally, while Layout-FID provides a robust way to assess structural and semantic alignment with seed samples, it is currently used in a post-hoc manner. A natural next step is to close the loop by integrating this metric into the generation pipeline enabling real-time scoring and iterative optimization during generation. For instance, low Layout-FID could be used as a reward signal in a reinforcement learning setup to promote higher fidelity outputs.

Similarly, the visual realism study shows that MLLM-guided distortions tend to hallucinate artifacts on cleaner domains, which elevates Layout-FID scores. This suggests an avenue to use Layout-FID not just for evaluation, but also as a diagnostic tool to guide selective realism application i.e., apply realism enhancements only when they yield meaningful improvements.

Overall, these findings motivate the development of:

- Diversity-aware generation modules that reduce redundancy across LLM calls.
- Feedback loops that use document-aware metrics to inform and refine generation.
- Realism modules that dynamically adapt based on the structure and noise profile of seed documents.
- Optimization strategies that integrate quality metrics as part of a training or fine-tuning signal.

Such advances are essential to evolve DocGenie from a generation-evaluation framework to a self-improving document synthesis system, making it more aligned with the needs of real-world document AI pipelines.

6. Conclusion

This work introduced **DocGenie**, a scalable framework for generating high-fidelity synthetic business documents to support document understanding tasks such as OCR, information extraction, and layout analysis. Unlike prior efforts that focus primarily on academic or structured research article datasets, DocGenie targets real-world business documents including invoices, receipts, forms, and budgets which are often semi-structured, visually diverse, and domain-specific. By leveraging a frontier multimodal LLM and seed-guided generation, DocGenie produces structured HTML documents that align with the linguistic, cultural, and visual characteristics of real business domains. A document-specific adaptation of the FID metric Layout-FID is proposed, using LayoutLMv3 embeddings to more effectively capture semantic, structural, and layout similarity than traditional image-based alternatives. Experiments across five business document categories demonstrate the framework's generalizability and document alignment capabilities. A comparative study of realism strategies highlights the strengths of human-guided enhancements and the limitations of current LLM automated approaches. DocGenie establishes a practical and extensible foundation for scalable synthetic document generation in business settings. Future work will explore feedback-aware generation, adaptive realism strategies, and expansion to multilingual and multimodal document domains.

7. Acknowledgments

The authors would like to thank Gourab Kumar Patro, Gopalakrishnan Saisubramaniam and Varun V for their valuable time and constructive feedback throughout the development of this work. Their detailed reviews and suggestions significantly contributed to improving the quality and clarity of the manuscript.

References

[1] wkhtmltopdf: Convert html to pdf using webkit. https://wkhtmltopdf.org/. Accessed: 2025-03-26. 5

[2] Anthropic. Claude 3.7 sonnet, 2025. Accessed: 2025-03-25. 5

[3] Sanket Biswas, Pau Riba, Josep Lladós, and Umapada Pal. Docsynth: a layout guided approach for controllable document image synthesis. In *International Conference on Document Analysis and Recognition*, pages 555–568. Springer, 2021. 1, 2

[4] Sanket Biswas, Rajiv Jain, Vlad I Morariu, Jiuxiang Gu, Puneet Mathur, Curtis Wigington, Tong Sun, and Josep Lladós. Docsynthv2: A practical autoregressive modeling for document generation. *arXiv preprint arXiv:2406.08354*, 2024. 1, 2

[5] Quang Anh Bui, David Mollard, and Salvatore Tabbone. Automatic synthetic document image generation using generative adversarial networks: application in mobile-captured document analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 393–400. IEEE, 2019. 2

[6] Chuanghao Ding, Xuejing Liu, Wei Tang, Juan Li, Xiaoliang Wang, Rui Zhao, Cam-Tu Nguyen, and Fei Tan. Synthdoc:

Bilingual documents synthesis for visual document understanding. In *Proceedings of the 2nd Workshop on Large Generative Models Meet Multimodal Applications*, pages 16–25, 2024. 2

[7] Tahani Fennir, Bart Lamiroy, and Jean-Charles Lamirel. Using gans for domain adaptive high resolution synthetic document generation. In *International Conference on Document Analysis and Recognition*, pages 49–61. Springer, 2023. 2

[8] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE, 2015. 1, 4

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1, 2

[10] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 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, 2022. 1, 2, 3, 5

[11] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019. 4

[12] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Layoutdm: Discrete diffusion model for controllable layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10167–10176, 2023. 2

[13] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Towards flexible multi-modal document models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14287–14296, 2023. 2

[14] Document IU. The competition on visually rich document intelligence and understanding (vrd-iu), 2025. 5

[15] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024. 2

[16] Nicholas Journet, Muriel Visani, Boris Mansencal, Kieu Van-Cuong, and Antoine Billy. Doccreator: A new software for creating synthetic ground-truthed document images. *Journal of imaging*, 3(4):62, 2017. 2

[17] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 1, 2, 4

[18] Marek Kozłowski and Paweł Weichbroth. Samples of electronic invoices, 2021. 5

[19] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3530–3539, 2022. 1

[20] Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. Geolayoutlm: Geometric pre-training for visual information extraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7092–7101, 2023. 1

[21] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15630–15640, 2024. 3

[22] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019. 4

[23] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 1

[24] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018. 1, 3

[25] Jaejung Seol, Seojun Kim, and Jaejun Yoo. Posterllama: Bridging design ability of language model to content-aware layout generation. In *European Conference on Computer Vision*, pages 451–468. Springer, 2024. 2

[26] Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalickỳ, Jiří Matas, Antoine Doucet, Mickaël Coustaty, et al. Docile benchmark for document information localization and extraction. In *International Conference on Document Analysis and Recognition*, pages 147–166. Springer, 2023. 2

[27] Zecheng Tang, Chenfei Wu, Juntao Li, and Nan Duan. Layoutnuwa: Revealing the hidden layout expertise of large language models. *arXiv preprint arXiv:2309.09506*, 2023. 2

[28] Noman Tanveer, Adnan Ul-Hasan, and Faisal Shafait. Diffusion models for document image generation. In *International Conference on Document Analysis and Recognition*, pages 438–453. Springer, 2023. 1, 3

[29] Yilin Wang, Zeyuan Chen, Liangjun Zhong, Zheng Ding, and Zhuowen Tu. Dolfin: Diffusion layout transformers without autoencoder. In *European Conference on Computer Vision*, pages 326–343. Springer, 2024. 2

[30] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 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*, pages 1192–1200, 2020. 1

[31] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1015–1022. IEEE, 2019. 1, 2

DocGenie: A Framework for High-Fidelity Synthetic Document Generation via Seed-Guided Multimodal LLM and Document-Aware Evaluation

Supplementary Material

8. Prompt Templates for Document Generation

To clarify the prompting strategies used for synthetic document creation, this section presents the exact templates employed during seed-free and seed-guided generation using Claude-Sonnet 3.7. These prompts reflect how the MLLM was conditioned to produce HTML-based business documents across various categories. The seed-guided version includes additional context in the form of real document examples, enabling stronger domain alignment and structural consistency. Both prompt variants were designed to elicit diverse yet semantically coherent outputs aligned with business document standards.

8.1. Seed-Free Generation Prompt

You are an AI specialized in generating multiple unique HTML documents in one response. Please create {num_solutions} unique HTML documents representing {doc_type}.
Each solution must:

1. Include all mandatory fields: {sections}.
2. Be formatted so it could print on A4 (e.g., use @page {{ size: A4; }} in your CSS).
3. Show a significantly different layout, styling, and textual content from every other solution.
4. Maintain a {background_requirements}.
5. Avoid copy-pasting or reusing large chunks of HTML, CSS, or disclaimers—each document must be at least 70% different in code and text than the others.
6. Wrap each complete document between <HTML> and </HTML> tags, labeled as:
   1. <HTML>...Solution #1...</HTML>
   2. <HTML>...Solution #2...</HTML>
   ...
   {num_solutions}. <HTML>...Solution #{num_solutions}...</HTML>

Do not provide additional commentary or references to the other solutions within each HTML.
Now generate the {num_solutions} distinct {doc_type} documents.

8.2. Seed-Guided Generation Prompt

You are an AI specialized in generating unique HTML documents based on multiple scanned images of real-world examples. You have been provided with distinct sample images, each from a different cultural or regional background. You have been provided seed images of {doc_type}, each originating from different cultural or regional contexts. For example, some might feature:

- Local languages or regional disclaimers
- Different date formats (e.g., dd/mm/yyyy vs. mm/dd/yyyy)
- Unique currency or numbering formats
- Varying layout norms (positions of key fields, disclaimers, official stamps, etc.)

Now, please generate {num_solutions} unique HTML documents that:

1. Strictly reflect the overall style, layout, and cultural cues found in these samples, but do NOT copy any text, disclaimers, or layout verbatim from the samples.
2. Include any essential mandatory fields: {sections}.
3. Maintain an A4 size format for printing (using @page {{ size: A4; }} or similar CSS).
4. Maintain a {background_requirements}.
5. Avoid copy-pasting or reusing large chunks of HTML, CSS, or disclaimers—each document must be at least 70% different in code and text than the others.
6. Strictly wrap each new document in <HTML>...</HTML> tags, for example:
   1. <HTML>...Solution #1...</HTML>
   2. <HTML>...Solution #2...</HTML>
   ...
   {num_solutions}. <HTML>...Solution #{num_solutions}...</HTML>

Additional Requirements: {user_descriptions}
Notes:

- Pay close attention to cultural/regional differences seen in the seed images (e.g., language, format, disclaimers).
- Feel free to creatively adapt or combine stylistic cues from the seeds, as long as the end result looks authentic for that cultural context.
- Do NOT directly copy-paste text or entire code blocks from any single seed image or across these new solutions.

Now please generate the {num_solutions} distinct {doc_type} documents.

9. Visual Examples of Realism Enhancement

To complement the quantitative Layout-FID analysis, this section presents qualitative comparisons of synthetic docu-

ments before and after realism enhancement across all five document categories. Each column in Figure 3 corresponds to a document type: CORD, SROIE, Invoice, Form604, and Budget. Four versions of each sample are shown in rows:

- **Real Seed Sample**: A real seed sample, used as a reference.
- **Raw (No Realism)**: Synthetic document generated by DocGenie without any post-processing.
- **Human-Guided Realism**: Post-processed using distortions selected through manual inspection of seed documents, based on common artifacts identified in the Donut paper.
- **MLLM-Guided Realism**: Post-processed using distortions dynamically predicted by Claude-Sonnet 3.7, guided by visual cues from seed documents.

Both the human and MLLM-guided approaches are effective at identifying realism components present in noisy document categories like CORD, where distortions are visually prominent in the seed samples. However, while the human-guided method continues to perform reliably across cleaner domains such as Form604, Invoice, and Budget, the MLLM-guided strategy tends to hallucinate distortions not present in the originals. This leads to the introduction of unrealistic artifacts such as perspective warping in Form604, excessive blur in Invoice, and artificial shadows or noise in Budget, ultimately reducing alignment with the seed distribution This supports the quantitative findings discussed in Section 4.3 and reinforces the importance of distortion-aware realism pipelines tailored to domain characteristics.

10. Repetition in MLLM-Based Document Generation

Due to the context length limitation of Claude-Sonnet 3.7 (8192 tokens), DocGenie generates a maximum of four HTML-based documents per LLM call. To construct a full set of ten synthetic samples per document category, this process is repeated across multiple independent calls. However, since the MLLM lacks memory of previously generated outputs, it may occasionally reproduce similar layouts or field structures across calls—particularly in domains with semi-structured or repetitive patterns.

This section presents qualitative examples illustrating such layout-level repetition in receipt and budget documents. In both cases, some synthetic samples generated in separate LLM calls exhibit near-identical layout structures and content flow. While not a pervasive issue across all document types, this highlights a potential limitation in generating highly diverse samples at scale without tracking prior generations.

A future enhancement could involve incorporating feedback loops that feed earlier generations back into the LLM context to encourage greater inter-sample variation.

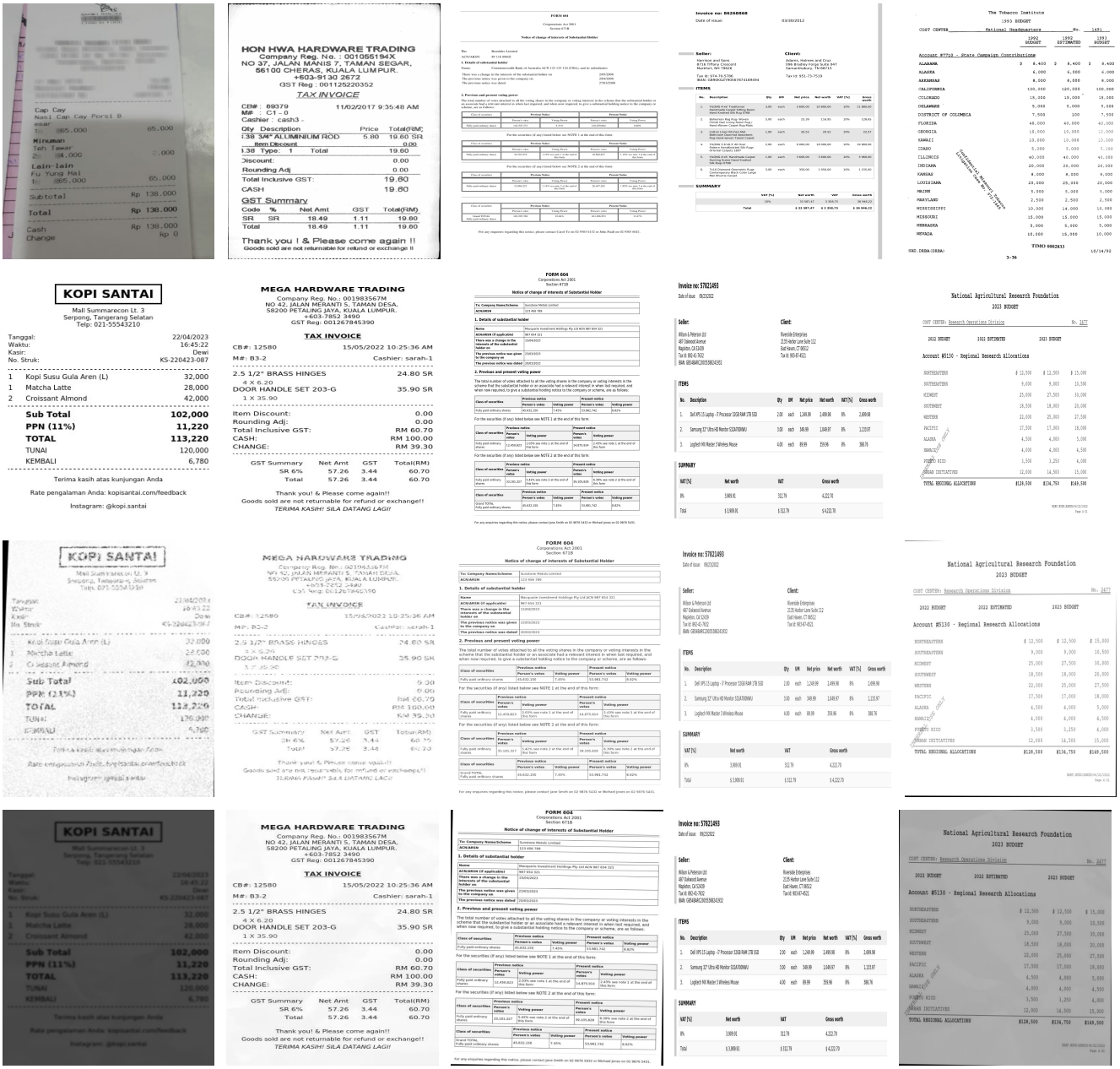
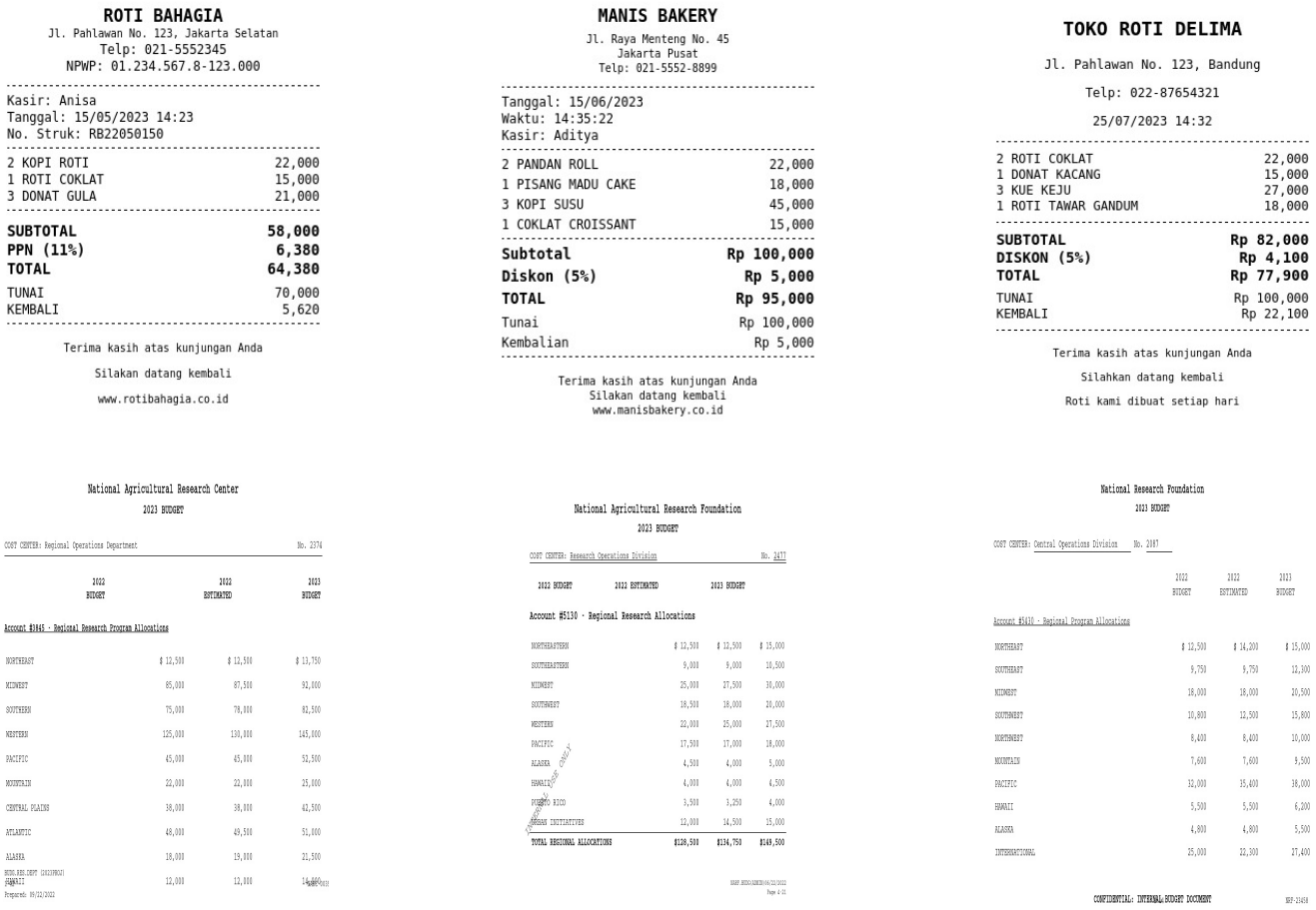


Figure 3. Visual comparison of realism enhancement strategies across document types. Each column shows a different document category: CORD, SROIE, Form604, Invoice, and Budget. Rows represent: (1) real seed samples, (2) synthetic documents without realism enhancement, (3) human-guided realism, and (4) MLLM-guided realism. Seed samples serve as visual references. Both human and MLLM-guided strategies accurately identify realism components in noisy domains like CORD and SROIE. However, human-guided realism remains consistent across cleaner domains, while the MLLM-guided method tends to hallucinate artifacts introducing exaggerated perspective warping in Form604, excessive blur in Invoice, and artificial shadows and Gaussian noise in Budget.



ROTI BAHAGIA
Jl. Pahlawan No. 123, Jakarta Selatan
Telp: 021-5552345
NPWP: 01.234.567.8-123.000

Kasir: Anisa
Tanggal: 15/05/2023 14:23
No. Struk: RB22050150

2 KOPI ROTI	22,000
1 ROTI COKLAT	15,000
3 DONAT GULA	21,000
SUBTOTAL	**58,000**
PPN (11%)	**6,380**
TOTAL	**64,380**
TUNAI	70,000
KEMBALI	5,620

Terima kasih atas kunjungan Anda
Silakan datang kembali
www.rotibahagia.co.id

MANIS BAKERY
Jl. Raya Menteng No. 45
Jakarta Pusat
Telp: 021-5552-8899

Tanggal: 15/06/2023
Waktu: 14:35:22
Kasir: Aditya

2 PANDAN ROLL	22,000
1 PISANG MADU CAKE	18,000
3 KOPI SUSU	45,000
1 COKLAT CROISSANT	15,000
Subtotal	**Rp 100,000**
Diskon (5%)	**Rp 5,000**
TOTAL	**Rp 95,000**
Tunai	Rp 100,000
Kembalian	Rp 5,000

Terima kasih atas kunjungan Anda
Silakan datang kembali
www.manisbakery.co.id

TOKO ROTI DELIMA
Jl. Pahlawan No. 123, Bandung
Telp: 022-87654321
25/07/2023 14:32

2 ROTI COKLAT	22,000
1 DONAT KACANG	15,000
3 KUE KEJU	27,000
1 ROTI TAWAR GANDUM	18,000
SUBTOTAL	**Rp 82,000**
DISKON (5%)	**Rp 4,100**
TOTAL	**Rp 77,900**
TUNAI	Rp 100,000
KEMBALI	Rp 22,100

Terima kasih atas kunjungan Anda
Silahkan datang kembali
Roti kami dibuat setiap hari

National Agricultural Research Center
2023 BUDGET
COST CENTER: Regional Operations Department — No. 2374

	2022 BUDGET	2022 ESTIMATED	2023 BUDGET
Account #3845 - Regional Research Program Allocations			
NORTHEAST	$ 12,500	$ 12,500	$ 13,750
MIDWEST	85,000	87,500	92,000
SOUTHERN	75,000	78,000	82,500
WESTERN	125,000	130,000	145,000
PACIFIC	45,000	45,000	52,500
MOUNTAIN	22,000	22,000	25,000
CENTRAL PLAINS	38,000	38,000	42,500
ATLANTIC	48,000	49,500	51,000
ALASKA	18,000	19,000	21,500
HAWAII	12,000	12,000	14,000

Prepared: 09/22/2022

National Agricultural Research Foundation
2023 BUDGET
COST CENTER: Research Operations Division — No. 2417

	2022 BUDGET	2022 ESTIMATED	2023 BUDGET
Account #5130 - Regional Research Allocations			
NORTHEASTERN	$ 12,500	$ 12,500	$ 15,000
SOUTHEASTERN	9,000	9,000	10,500
MIDWEST	25,000	27,500	30,000
SOUTHWEST	18,500	18,000	20,000
WESTERN	22,000	25,000	27,500
PACIFIC	17,500	17,000	18,000
ALASKA	4,500	4,000	5,000
HAWAII	4,000	4,000	4,500
PUERTO RICO	3,500	3,250	4,000
URBAN INITIATIVES	12,000	14,500	15,000
TOTAL REGIONAL ALLOCATIONS	$120,500	$134,750	$149,500

Page 4 of 11

National Research Foundation
2023 BUDGET
COST CENTER: Central Operations Division — No. 2307

	2022 BUDGET	2022 ESTIMATED	2023 BUDGET
Account #5420 - Regional Program Allocations			
NORTHEAST	$ 12,500	$ 14,200	$ 15,000
SOUTHEAST	9,750	9,750	12,300
MIDWEST	18,000	18,000	20,500
SOUTHWEST	10,800	12,500	15,800
NORTHWEST	8,400	8,400	10,000
MOUNTAIN	7,600	7,600	9,500
PACIFIC	32,000	35,400	38,000
HAWAII	5,500	5,500	6,200
ALASKA	4,800	4,800	5,500
INTERNATIONAL	25,000	22,300	27,400

CONFIDENTIAL: INTERNAL BUDGET DOCUMENT

Figure 4. Layout repetition observed in the first sample generated from each of three separate LLM calls (corresponding to indices 1, 5, and 9) for two document categories: CORD receipts and Budget documents. Although the documents were generated independently across iterations, their layouts and content structure appear highly similar. This illustrates a potential limitation of DocGenie without memory of prior outputs.