VITA: Visual In-image Text Analysis on Vision-Language Models

Anonymous ACL submission

Abstract

Vision-language models (VLMs) excel at document summarization, yet their robustness to visual formatting variations in document images is not well understood. We present VITA (Visual In-image Text Analysis), a systematic framework that measures how realistic visual changes—text emphasis and structural formatting—affect summarization quality. We evaluate six VLMs spanning early, middle, and late fusion architectures at two model scales. Across lexical, semantic, and information preservation metrics, we find architecture- and scale-dependent vulnerabilities: early fusion loses more information despite higher lexical stability, whereas late fusion preserves information but exhibits larger lexical variation. Structural formatting induces larger degradations than text emphasis. Scaling mitigates emphasis sensitivity but can exacerbate structural-format vulnerabilities in late fusion, indicating that robust document understanding may require architectural innovations beyond scaling. Source code and dataset are available at [Under Review]

1 Introduction

Vision-language models (VLMs) integrate visual and textual modalities for cross-modal reasoning (Zhang et al., 2024a; Chen et al., 2024a; Gigant et al., 2025; Orna et al., 2024; G et al., 2024). With advances in large-scale pretraining and fusion architectures, VLMs have expanded beyond classic multimodal tasks (Achiam et al., 2023; Dai et al., 2023; Li et al., 2023) to document understanding and summarization, where inputs are rendered document images (e.g., screenshots and PDFs) containing both text and visual cues (Nacson et al., 2025; Lee et al., 2023).

Unlike OCR-centric pipelines, VLMs can process document images directly (Faysse et al., 2024), leveraging visual signals (e.g., font, color, alignment, spatial layout) alongside textual semantics. While such cues can convey emphasis, hierarchy, and structure, they can also induce brittleness: models may overfit to superficial styles, behave inconsistently across layouts/domains, or rely on non-semantic artifacts. This motivates the following research questions:

- How strongly do visual presentation cues affect the quality of VLM-generated document summaries?
- Which visual cues are most likely to change outputs even when the underlying text is identical?
- Do VLMs leverage meaningful visual signals, or are they overly sensitive to superficial stylistic variations that harm generalization?

A second key factor is *how* VLMs fuse visual and textual information (Hemker et al., 2024; Zhang et al., 2024b; Gavrikov et al., 2025). Fusion may occur early (token concatenation), in the middle (cross-attention), or late (representation-level merging), potentially shaping how visual cues influence generation. This raises an additional question:

- Does fusion timing and mechanism shape how VLMs process and respond to visual presentation cues?

Finally, a critical practical consideration is whether increased model capacity can mitigate these sensitivities. Scaling laws suggest that larger models generally achieve better performance (Kaplan et al., 2020), but recent work has shown that robustness does not automatically improve with scale (Howe et al., 2024), and some tasks even exhibit inverse scaling where larger models perform worse (McKenzie et al., 2023). This raises a final question:

- Does increased model scale reduce sensitivity to visual presentation cues, and if so, does this

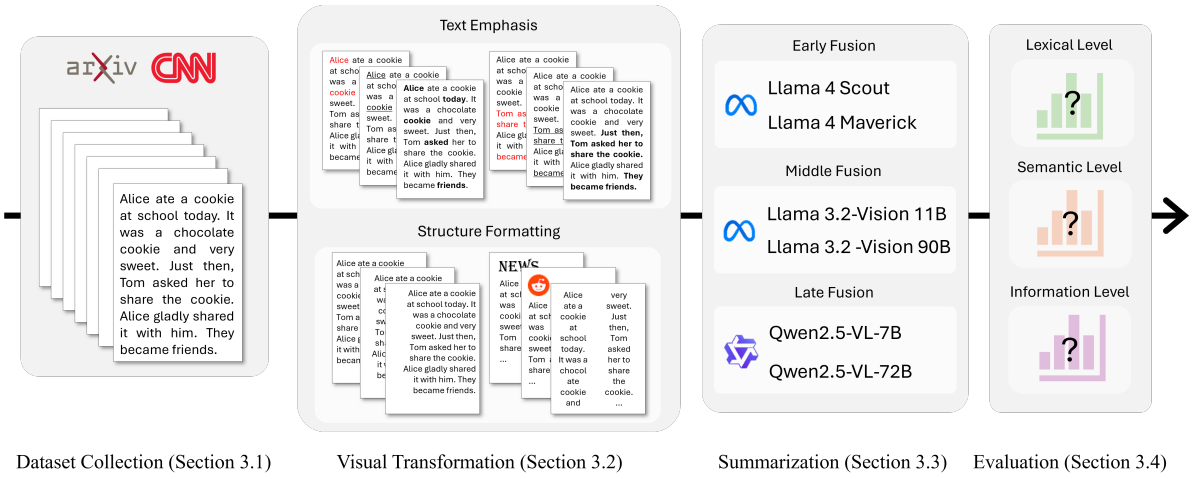

Figure 1: Overview of VITA. We keep the underlying text fixed and vary only visual attributes to form paired document images (text emphasis vs. structural formatting). We evaluate six VLMs across fusion timing (early/middle/late) and model scale, and quantify summary distortion using lexical/semantic similarity, verification-based faithfulness, and the rate of transformation-specific clue words that indicate format-aware summarization.

improvement generalize across all transformation types?

These questions matter because real-world documents vary widely in presentation (e.g., reports/news vs. social media/emails), and sensitivity to superficial formatting can undermine reliability. Yet controlled evidence isolating which visual attributes drive summarization changes remains scarce: prior work often evaluates document understanding without holding text constant under visual variations (Mathew et al., 2021; Ouyang et al., 2025; Qiu et al., 2022), and comparative analyses across fusion strategies and model scales are limited.

To fill this gap, we introduce **VITA (Visual In-image Text Analysis)**, a framework that isolates and quantifies the impact of realistic visual text variations on VLM-based summarization (Figure 1). We construct paired-input experiments with fixed text and varied visual attributes, organized into (1) text emphasis (e.g., color, bold, underline) and (2) structure formatting (e.g., alignment, columns, platform-style templates). We evaluate outputs via lexical drift (TF-IDF cosine similarity (Salton, 1983)), semantic change (SentenceBERT (Reimers and Gurevych, 2019)), and key information preservation (Minicheck (Tang et al., 2024)). We benchmark six representative VLMs spanning early, middle, and late fusion at two scales: Llama 4 Scout/Maverick (early), Llama 3.2-Vision 11B/90B (middle), and Qwen2.5-VL-7B/72B (late).

Our main contributions are summarized as follows:

- We propose VITA, a systematic analysis framework that uses controlled paired-input experiments to isolate visual changes while holding text constant.
- We analyze six VLMs across early, middle, and late fusion at two model scales, examining how fusion timing and capacity modulate sensitivity to visual presentation cues.
- We release a benchmark dataset of 2,600 document images with 12 visual transformations to facilitate future research.
- We provide quantitative and qualitative analyses revealing that (i) structure formatting causes greater degradation than text emphasis, (ii) fusion architectures exhibit distinct vulnerability patterns, and (iii) scaling selectively mitigates text emphasis sensitivity while leaving design transformation vulnerabilities unresolved.

2 Related Work

2.1 Visual Text Analysis

While substantial progress has been made in evaluating VLM performance on document understanding tasks, existing benchmarks exhibit significant limitations in systematically analyzing visual text variations. Traditional document VQA benchmarks such as DocVQA (Mathew et al., 2021), OCRBench (Fu et al., 2024), and OmniDocBench

(Ouyang et al., 2025) primarily utilize naturally occurring document images with diverse visual presentations. However, these datasets lack controlled experimental designs that enable direct comparison between identical textual content under different visual conditions, preventing systematic investigation of how specific visual attributes influence model behavior.

Recent studies have attempted to address this gap through more controlled approaches. Verma et al. (2023) introduced cross-lingual document pairs with identical content, while Chen et al. (2024b) evaluated VLM robustness by applying systematic transformations to document images. However, these approaches focus on linguistic variations or image-level degradations rather than fine-grained textual visual attributes commonly encountered in real-world documents.

A critical gap persists in prior work: the absence of systematic frameworks for isolating and quantifying the effects of specific visual text attributes on VLM performance. Current approaches lack the granular control necessary to understand how individual visual elements—such as text emphasis and structural formatting—influence model outputs when textual content remains constant.

2.2 Multimodal Fusion Structure

Recent studies have shown that the timing of multimodal fusion significantly affects model performance and sensitivity to external factors (Hemker et al., 2024; Zhang et al., 2024b; Gavrikov et al., 2025). Hemker et al. (2024) improved robustness to missing modalities by combining early fusion with attention mechanisms, Zhang et al. (2024b) achieved superior performance in document layout analysis through early-late fusion combinations, and Gavrikov et al. (2025) demonstrated that multimodal fusion itself changes how visual cues are processed. Additionally, Qiu et al. (2022) confirmed that modifications to the textual content cause the most severe distributional shifts in multimodal models when partially replacing text within images. However, existing studies have focused on general performance comparisons (Mathew et al., 2021; Fu et al., 2024; Ouyang et al., 2025). Despite evidence that fusion timing affects visual feature processing, no study has systematically analyzed how visual text attribute transformations impact different fusion architectures while keeping textual content constant.

2.3 Scaling and Robustness

Scaling laws have established that language model performance improves predictably with increased model size, training data, and compute (Kaplan et al., 2020). This relationship has driven the development of increasingly large models across both language and vision-language domains. However, the relationship between scale and robustness is more nuanced than general performance metrics suggest.

Howe et al. (2024) found that larger models are not consistently more robust to adversarial attacks without explicit safety training. More critically, McKenzie et al. (2023) demonstrated inverse scaling on 11 tasks where larger models performed worse than smaller ones, attributing this to memorization preferences and imitation of undesirable training patterns. In the multimodal domain, whether scaling mitigates sensitivity to visual presentation cues in document understanding remains unexplored—a significant gap given that VLMs must handle interactions between visual and textual modalities where scaling effects may differ from text-only models.

3 VITA: Visual In-image Text Analysis

To analyze how visual in-image text affects VLM-based document summarization, we propose VITA, short for Visual In-image Text Analysis. VITA isolates visual factors while keeping textual content constant, enabling controlled experiments over realistic visual variations and comparisons across fusion strategies. As shown in Figure 1, VITA consists of four stages: dataset collection, visual transformation, summarization, and evaluation.

3.1 Dataset Collection

We build a domain-diverse summarization dataset from two common sources: scientific papers and news articles. We collected 100 documents each from arXiv and CNN, restricting the time window to March–May 2025 to reduce potential overlap with VLM pretraining data [1]. For arXiv papers, we extract only the introduction; for news, we use the full article body. Each text is rendered into a single document image using LaTeX.

[1]We will release a publicly available code for data collection.

(a) Text Emphasis

		Color		Bold		Underline				Color		Bold		Underline	
		Word	Sent	Word	Sent	Word	Sent			Word	Sent	Word	Sent	Word	Sent
Early Fusion								*Early Fusion*							
Llama-4 Scout	LEX	-.00	.00	.01	.01	.01*	.01	Llama-4 Maverick	LEX	.01	.01	.00	-.00	.01	.00
	SEM	-.00	-.01	-.01*	-.01*	-.01	-.01		SEM	-.01	-.01*	-.00	-.02**	-.01	-.00
	INFO	**.20*****	**.18*****	**.19*****	**.17*****	**.19*****	**.18*****		INFO	.00	-.02	.01	-.03	-.03	-.01
Middle Fusion								*Middle Fusion*							
Llama-3.2 11B	LEX	.06*	.04	.02	.04*	.03	.04*	Llama-3.2 90B	LEX	-.01	-.01	-.00	-.02	.01	-.00
	SEM	.03*	.02	.01	.02	.02*	.03*		SEM	-.00	-.00	.01	-.01	.00	-.01
	INFO	**.15*****	**.14*****	**.11*****	**.12*****	**.11*****	**.11*****		INFO	-.02	-.01	-.01	-.03*	-.00	-.01
Late Fusion								*Late Fusion*							
Qwen2.5 7B	LEX	.00	.03***	.02**	.05***	.04***	.04***	Qwen2.5 72B	LEX	-.00	-.01	.01	.01	.01	.01*
	SEM	.01	.02*	.01	.01***	.02***	.01*		SEM	-.00	-.00	-.00	.00	-.00	.00
	INFO	.05*	.06	.00	.00	.03	-.01		INFO	.00	.01	-.00	-.01	.01	-.00

(b) Structure Formatting

		Alignment			Col	Design				Alignment			Col	Design	
		Left	Cen	Right	Two	Reddit	News			Left	Cen	Right	Two	Reddit	News
Early Fusion								*Early Fusion*							
Llama-4 Scout	LEX	.00	-.00	.01	.01	.06***	.03	Llama-4 Maverick	LEX	-.00	-.01	-.01	.01	.04***	.03*
	SEM	-.00	-.00	.02	.00	.06***	.04**		SEM	.00	.01	.00	.02**	**.10*****	.09***
	INFO	**.20*****	**.16*****	**.18*****	**.21*****	**.23*****	**.47*****		INFO	.01	-.00	.01	.06***	**.34*****	**.39*****
Middle Fusion								*Middle Fusion*							
Llama-3.2 11B	LEX	-.00	.06	.05*	.06*	**.15*****	.09***	Llama-3.2 90B	LEX	-.01	-.00	-.01	.01	.05***	.02
	SEM	.02	.06***	.03	.05*	**.16*****	.06***		SEM	-.00	.01*	.00	.02*	**.10*****	**.10*****
	INFO	**.10*****	**.13*****	**.12*****	**.15*****	**.22*****	**.25*****		INFO	-.01	-.00	-.01	.07***	**.36*****	**.40*****
Late Fusion								*Late Fusion*							
Qwen2.5 7B	LEX	.02***	.01	.02	.02*	.05***	.06***	Qwen2.5 72B	LEX	-.00	.00	.00	.02	.02***	.03***
	SEM	.01	.01	.01	.01*	.02***	.02***		SEM	-.01*	-.01**	.00	.01	.01**	.00
	INFO	.03	.02	.01	.01	.01	.01		INFO	.00	.00	-.00	.01	.04**	**.26*****

$^{*}\,p < 0.05$, $^{**}\,p < 0.01$, $^{***}\,p < 0.001$

Table 1: Performance degradation of six VLMs across 12 visual text transformations. LEX/SEM/INFO denote lexical, semantic, and information-level changes. Cen: center; Col Two: two-column. Left: small-scale; Right: large-scale models. Bold indicates values ≥ 0.10.

3.2 Visual Transformation

Each document is programmatically styled in LaTeX to create 12 transformed versions, then compiled into images. The transformations mirror real-world document variations and fall into two groups: text emphasis, which highlights salient content, and structure formatting, which alters global layout and design.

Text Emphasis. We apply six transformations at both word and sentence levels, using color, bold, and underline. To select emphasis targets, we run gradient-based saliency analysis and highlight the top 15% of words that most influence model outputs. Goyal et al. (2022); Narayan et al. (2018) show that the top 10–20% saliency regions align well with key summarization content, and Hollenstein and Beinborn (2021) report partial alignment with human importance judgements.

Structure Formatting. We study global layout effects with three transformations of increasing intensity. We first vary alignment across left, center, and right. We then apply a two-column layout, and finally reformat the same content into two distinct designs, namely a newspaper article and a Reddit post. All designs except Reddit are implemented in LaTeX. Reddit is rendered via HTML and captured as an image.

3.3 Summarization

We evaluate six VLMs spanning early, middle, and late fusion at both small and large scales to test how fusion depth and capacity modulate sensitivity to visual cues. We follow a standardized protocol and prompt each model with *"Please summarize the text in this image."* We generate summaries for the original image and 12 transformed images, yielding 2,600 image-summary pairs from 200 base documents under 13 conditions, evaluated across two model scales for each fusion architecture.

Early Fusion. We use Llama 4 Scout with 109B parameters and Llama 4 Maverick with 402B parameters as early-fusion models. These models combine vision features with text tokens and process them jointly in a single backbone (Meta, 2024). This tight integration can strongly propagate fine-grained cues, such as color, font, and layout, into semantic representations.

Middle Fusion. LLaMA 3.2-Vision 11B and 90B represent middle fusion, where vision features interact with text via cross-attention at intermediate layers (Dubey et al., 2024). This design may partially transmit document-level visual transformations into summaries.

Late Fusion. We adopt Qwen2.5-VL-7B and Qwen2.5-VL-72B as late-fusion models. These models largely process visual information separately and merge it with text representations at later stages (Bai et al., 2025), which can limit the influence of fine-grained visual styles on summarization.

3.4 Evaluation

We compare summaries from original and transformed images using three complementary dimensions: lexical, semantic, and informational quality. For each transformation, we conduct paired comparisons and test score-difference normality; we apply paired t-tests (Hsu and Lachenbruch, 2014) when normal, and Wilcoxon signed-rank tests (Woolson, 2007) otherwise.

Lexical. We measure lexical similarity as TF-IDF cosine similarity between summaries from the original and transformed images:

$$\text{Lex}(S_o, S_t) = \cos\left(\text{TF}(S_o), \text{TF}(S_t)\right) \quad (1)$$

where S_o and S_t denote summaries from the original and transformed images, and $\text{TF}(\cdot)$ denotes the TF-IDF vector. Lower scores indicate larger vocabulary shifts.

Semantic. We compute cosine similarity between Sentence-BERT embeddings (Reimers and Gurevych, 2019):

$$\text{Sem}(S_o, S_t) = \cos\left(\mathbf{e}(S_o), \mathbf{e}(S_t)\right) \quad (2)$$

where $\mathbf{e}(\cdot)$ denotes the SBERT embedding.

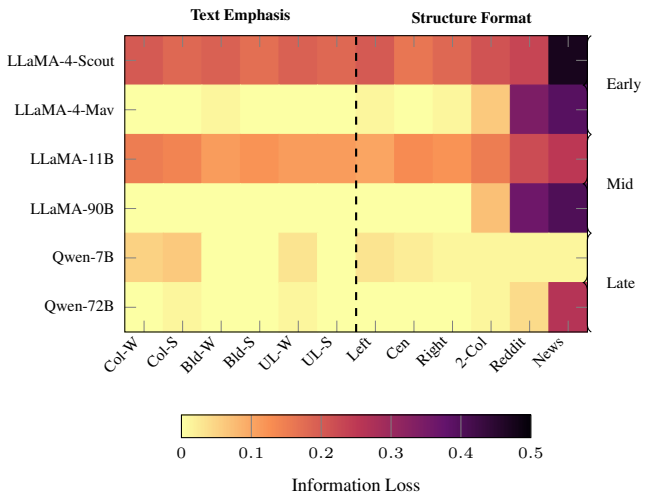

Figure 2: Information loss heatmap. Darker = higher degradation. Late Fusion (Qwen) is most robust; Design transformations cause severe degradation in Early/Middle Fusion.

Information. We evaluate factual consistency using Bespoke-MiniCheck-7B (Tang et al., 2024):

$$\text{Info}(S, D) = \text{MiniCheck}(S, D) \quad (3)$$

where D is the original document.

4 Results & Discussion

We analyze VITA results along three axes: (1) how models respond to different types of visual transformations, (2) how sensitivity patterns differ across fusion architectures, and (3) whether scaling mitigates these sensitivities uniformly. Table 1 presents performance changes for all six models, with small-scale models on the left and large-scale models on the right. Higher values indicate greater deviation from the original summary.

4.1 Visual Bias of In-Image Text

We first examine which types of visual transformations most strongly affect VLM summarization performance. Prior work suggests that document layout plays an important role in VLM processing (Zhang et al., 2024b), but systematic comparisons between local text modifications (e.g., color, bold) and global structural changes (e.g., layout, design templates) remain unexplored. Figure 2 visualizes information loss across all model-transformation combinations, revealing that Structure Formatting (right columns) causes substantially greater degradation than Text Emphasis (left columns), with Design transformations showing the most severe impact.

Text Emphasis	
Color	The generating function of the stabilized coefficients is ....involving the partition number into red colors.
Bold	The headline reads "Two Russian men ... The headline is written in large, bold font and is centered at the top of the page.
Underline	[NOT Detected]
Structure Formatting	
Alignment	It is aligned to the right side of the page...
Two Col	The article is divided into two columns... left column containing the introduction to the paper and the right column featuring..
News	The layout is typical of a newspaper article, with a clear heading, date, and body text...
Reddit	The image shows a screenshot of a Reddit page with a post about women in IT and software engineering.

Table 2: Examples of transformation-specific terminology in VLM outputs.

		LLaMA-4 Scout	LLaMA-3.2 11B	Qwen2.5 7B	LLaMA-4 Maverick	LLaMA-3.2 90B	Qwen2.5 72B
Text Emphasis							
Color	word	0.03	0.02	0.04	0.01	0.01	0.01
	sent	0.00	0.00	0.00	0.01	0.01	0.02
Bold	word	0.01	0.05	0.00	0.00	0.00	0.01
	sent	0.00	0.00	0.00	0.00	0.00	0.01
Underline	word	0.00	0.00	0.00	0.00	0.01	0.00
	sent	0.00	0.00	0.00	0.01	0.00	0.00
Structure Formatting							
Alignment	left	0.00	0.10	0.00	0.02	0.03	0.02
	center	0.04	0.09	0.03	0.05	0.04	0.04
	right	0.00	0.13	0.00	0.01	0.04	0.01
Column	two	0.03	**0.19**	0.00	**0.27**	**0.34**	0.00
Design	reddit	**0.42**	**1.00**	0.03	**0.90**	**0.99**	0.01
	news	**0.17**	**0.36**	0.03	**0.19**	**0.23**	0.04

Table 3: Proportion of summaries containing clue words for each visual condition.

Text Emphasis. Text Emphasis transformations demonstrate remarkable stability across all models and scales. LEXICAL changes remain minimal (0.00–0.06) and SEMANTIC alterations are equally constrained (0.00–0.03), indicating that visual emphasis largely preserves both vocabulary choice and meaning structure. Among transformation types, `Color` shows the highest stability with only 9 statistically significant effects across all conditions, compared to `Bold` (10) and `Underline` (12). This suggests that VLMs can largely filter out local text styling variations when processing document content.

Structure Formatting. In contrast, Structure Formatting transformations reveal a clear hierarchy of impact. `Alignment` and `Column` modifications show moderate effects (lexical 0.00–0.06, information loss 0.01–0.21), while `Design` transformations cause severe degradation across all metrics (lexical 0.03–0.15, semantic 0.02–0.16, information loss 0.01–0.47). Within Design transformations, `Reddit` format induces higher lexical change (0.05–0.15) than `News` format (0.03–0.09), though News shows comparable or higher information loss in several models.

Design as Critical Threshold. Qualitative analysis reveals that Design transformations trigger a fundamental shift in model processing. As shown in Table 2, Text Emphasis occasionally generates format-aware references (e.g., "involving the partition number into red colors") while preserving content focus. However, Design transformations cause complete processing mode shifts: News outputs include "The layout is typical of a newspaper article, with a clear heading, date, and body text..." while Reddit produces "The image shows a screenshot of a Reddit page with a post about...," abandoning content analysis entirely in favor of format identification.

The clue word analysis (Table 3) quantifies this phenomenon. Text Emphasis shows minimal format terminology incorporation (at most 0.05 across all models), while Design transformations trigger dramatic increases—Reddit format causes terminology usage in 42% (Early), 100% (Middle), and only 3% (Late) of summaries. These patterns

Model	*Contents*
Early	The image appears to be a mock-up of the front page of The New York Times... / The image shows a Reddit page with a post about quantum mechanics.
Middle	The image presents a page from ... featuring an article on quantum mechanics.. / The image shows a Reddit post... The post discusses the variational method..
Late	The text discusses the variational method in quantum mechanics, particularly for calculating the energy levels.. / The text discusses the variational method in quantum mechanics, particularly for calculating the energy levels..

Table 4: Examples of fusion architecture effects on design transformation.

demonstrate that VLMs are fundamentally more sensitive to global layout changes than to local text styling, with Design transformations serving as a critical threshold that triggers complete shifts from content understanding to format recognition. At larger scales, this pattern persists: scaling eliminates Text Emphasis sensitivity but leaves Design vulnerabilities largely unresolved (Section 4.3).

4.2 Effect of Vision-Language Fusion

We next investigate how fusion architecture modulates sensitivity to visual transformations. While prior work has shown that fusion timing affects visual feature processing (Gavrikov et al., 2025), its specific impact on document summarization under controlled visual variations is unknown. We examine whether different fusion strategies exhibit distinct vulnerability patterns.

Early Fusion. At small scale, Llama 4 Scout exhibits a distinctive pattern: minimal lexical and semantic changes (at most 0.01) but severe information loss (0.17–0.20) under Text Emphasis, representing the highest information degradation among the three architectures. Design transformations cause extreme vulnerability, with News format recording 0.47 information loss—the maximum observed in our study. This pattern likely stems from Early Fusion's architectural characteristics, where the vision encoder is deeply integrated into the language model backbone, causing layout changes to potentially disrupt the entire attention mechanism. At larger scale, Maverick eliminates Text Emphasis sensitivity but maintains Design vulnerability (Reddit: 0.34, News: 0.39), with clue word usage increasing from 0.42 to 0.90 for Reddit.

Middle Fusion. At small scale, Llama 3.2-Vision 11B shows a graduated response pattern: moderate lexical changes (0.02–0.06) alongside substantial information loss (0.11–0.15) for Text Emphasis, and vulnerability approaching Early Fusion for Design (0.22–0.25). The model shows the highest format awareness, with Reddit transformations triggering format-specific terminology in 100% of summaries (Table 3), suggesting that cross-attention mechanisms actively attend to visual formatting cues. The 90B variant achieves Text Emphasis robustness but shows increased Design sensitivity (Reddit: 0.36, News: 0.40), with near-saturated Reddit clue word detection (0.99).

Late Fusion. At small scale, Qwen2.5-VL-7B demonstrates a pattern opposite to Early Fusion: low information loss across all transformations (maximum 0.06), including Design changes (0.01), while showing slightly higher lexical variations (0.02–0.06). This robustness likely results from the MLP-based Vision-Language Merger architecture, which processes visual information largely separately before merging with text representations. Late Fusion maintains exceptional resistance to format-level distractions, with clue word usage below 4% even under Design transformations. As shown in Table **??**, while Early and Middle Fusion outputs explicitly reference the visual format, Late Fusion consistently focuses on content. However, at 72B, unexpected News vulnerability emerges (0.26) despite Reddit stability (0.04), revealing implicit format biases even in robust architectures. Taken together, each fusion architecture exhibits a distinct vulnerability profile: Early Fusion sacrifices information for lexical stability, Late Fusion shows the opposite pattern, and no architecture achieves comprehensive robustness.

4.3 Effect of Model Scale

Table 5 summarizes scaling effects on Design transformations. Figure 3 visualizes the contrasting scaling effects: Text Emphasis degradation drops to near-zero across all architectures (left), while Design vulnerabilities persist or worsen with scale (right).

For Text Emphasis, large-scale models generally exhibit near-zero Info-loss shifts across archi-

Fusion	Scale	Info Loss Reddit	Info Loss News	Δ Reddit	Δ News
Early Fusion					
Llama-4	Scout	0.23	0.47	+0.11	–0.08
	Maverick	0.34	0.39		
Middle Fusion					
Llama-3.2	11B	0.22	0.25	+0.14	+0.15
	90B	0.36	0.40		
Late Fusion					
Qwen2.5	7B	0.01	0.01	+0.03	**+0.25**
	72B	0.04	**0.26**		

Table 5: Fusion-Scale interaction on Design transformation. Positive Δ indicates increased vulnerability with scaling. Bold indicates values ≥ 0.25.

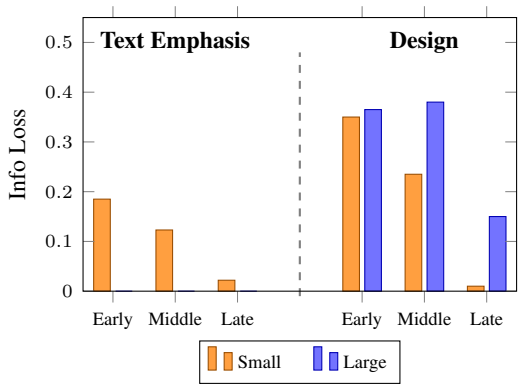

Figure 3: Scaling effect on information loss. Text Emphasis (left): scaling reduces degradation to near-zero. Design (right): scaling fails to mitigate vulnerability, with Middle Fusion showing inverse scaling.

tectures (Table 1a), suggesting reduced sensitivity with scaling. In contrast, Design transformations show architecture-dependent inverse-scaling tendencies (McKenzie et al., 2023): Early Fusion is mixed (Reddit +0.11, News –0.08), Middle Fusion consistently worsens (Reddit +0.14, News +0.15), and Late Fusion shows the sharpest divergence—Reddit remains stable (+0.03) while News increases from 0.01 to 0.26.

We hypothesize that larger models encode stronger format-specific priors, consistent with McKenzie et al.'s observation that scaling can amplify "imitation of undesirable training patterns." Notably, Late Fusion's News vulnerability emerges despite weak surface cue signals, suggesting an implicit format bias at deeper representational levels not captured by lexical diagnostics. Overall, scaling selectively mitigates visual sensitivity, and robust document understanding likely requires architectural changes beyond scaling alone.

5 Conclusion

This paper introduced VITA (Visual In-image Text Analysis), a systematic framework for analyzing how visual text transformations in document images affect VLM summarization. Through controlled paired-comparison experiments over 2,600 image–summary pairs and six VLMs spanning three fusion architectures at two scales, we addressed three research questions on visual sensitivity in document understanding.

First, we found that Structure Formatting transformations induce substantially larger degradation than Text Emphasis modifications. In particular, Design transformations (Reddit- and News-style layouts) act as a critical threshold: they reliably shift model behavior from content-focused summarization toward format-driven processing, producing consistent information loss across both design formats.

Second, fusion architectures exhibit distinct vulnerability profiles. Early Fusion tends to preserve lexical stability at the cost of information retention, whereas Late Fusion shows the opposite trade-off. Middle Fusion presents intermediate behavior with strong format awareness. Importantly, no single architecture provides comprehensive robustness across transformation families.

Third, we show that scaling selectively reduces visual sensitivity. Larger models generally attenuate vulnerabilities to Text Emphasis, suggesting that many emphasis cues are treated as shallow perturbations at higher capacity. However, sensitivity to Design transformations persists and can even increase with scale, consistent with prior observations of inverse-scaling tendencies where larger models amplify undesirable format priors. The Late Fusion contrast—where the 72B model exhibits emergent News vulnerability despite the 7B model remaining stable—highlights that scaling can introduce implicit format biases not captured by surface-level diagnostics.

Taken together, our findings caution against assuming that larger VLMs are inherently more robust to document format variation. Achieving reliable document understanding likely requires architectural and training interventions explicitly targeting format invariance, such as content–style disentanglement, format-agnostic objectives, or structure-aware calibration.

6 Limitations

Our results are based on statistically significant paired comparisons and show consistent Design effects across both Reddit- and News-style layouts, suggesting that the identified vulnerabilities are not driven by a single template. Nonetheless, our study has limitations.

First, our dataset size is moderate (200 documents per domain). Future work should validate the same trends at larger scale and across more diverse document genres and domains.

Second, we evaluate a controlled set of 12 transformations and six VLMs spanning three fusion architectures at two scales. Additional models (e.g., different instruction-tuning recipes or OCR-centric pipelines) and broader real-world formatting variations may reveal further nuances.

Third, VITA is primarily diagnostic: we focus on identifying "where and how" failures occur rather than proposing mitigations. Developing and testing interventions—such as format-robust augmentation, explicit content–style disentanglement, or architecture-level constraints for format invariance—remains important future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Yufan Chen, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ruiping Liu, Philip Torr, and Rainer Stiefelhagen. 2024b. Rodla: Benchmarking the robustness of document layout analysis models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15556–15566.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, and 1 others. 2024. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*.

Abinaya G, Guttulapavan Durga Rao, Pragada Sri Ram Gopal, Kaushik M, and B Sai Venkata Vignesh. 2024. Automated document processing: Combining ocr and generative ai for efficient text extraction and summarization. In *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–5.

Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. 2025. Can we talk models into seeing the world differently? In *Thirteenth International Conference on Learning Representations*. OpenReview. net.

Théo Gigant, Camille Guinaudeau, and Frédéric Dufaux. 2025. Summarization of multimodal presentations with vision-language models: Study of the effect of modalities and structure. *Preprint*, arXiv:2504.10049.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. 2024. Healnet: Multimodal fusion for heterogeneous biomedical data. *Advances in Neural Information Processing Systems*, 37:64479–64498.

Nora Hollenstein and Lisa Beinborn. 2021. Relative importance in sentence processing. *arXiv preprint arXiv:2106.03471*.

Nikolaus Howe, Ian McKenzie, Oskar Hollinsworth, Michał Zajac, Tom Tseng, Aaron Tucker, Pierre-Luc Bacon, and Adam Gleave. 2024. Scaling trends in language model robustness. *arXiv preprint arXiv:2407.18213*.

Henry Hsu and Peter A Lachenbruch. 2014. Paired t test. *Wiley StatsRef: statistics reference online*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, and 1 others. 2023. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*.

Meta. 2024. Introducing llama 4: Advancing multimodal intelligence.

Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. 2025. Docvlm: Make your vlm an efficient reader. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29005–29015.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Maliha Afroj Orna, Fariha Akther, and Md Abdul Masud. 2024. Ocr generated text summarization using bart. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 2128–2133.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, and 1 others. 2025. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24838–24848.

Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. 2022. Benchmarking robustness of multimodal image-text models under distribution shift. *arXiv preprint arXiv:2212.08044*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Gerard Salton. 1983. Modern information retrieval. *(No Title)*.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*.

Yash Verma, Anubhav Jangra, Raghvendra Kumar, and Sriparna Saha. 2023. Large scale multi-lingual multi-modal summarization dataset. *arXiv preprint arXiv:2302.06560*.

Robert F Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644.

Ning Zhang, Hiuyi Cheng, Jiayu Chen, Zongyuan Jiang, Jun Huang, Yang Xue, and Lianwen Jin. 2024b. M2doc: a multi-modal fusion approach for document layout analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7233–7241.

The Use of Large Language Models

We used AI assistance tools during the writing process of this manuscript. Specifically, we employed Grammarly for grammar checking, and GPT-5 for language polishing and improving clarity of expression. These tools were used for editorial purposes.

A Dataset Example

Figure 4 illustrates representative examples of the visual transformations used in our VITA framework.

B Detailed Results

We report per-model mean similarity scores (mean$\pm$std.) between summaries generated from the base image and summaries generated under each visual transformation. We measure three dimensions: lexical similarity (TF-IDF cosine similarity), semantic similarity (Sentence-BERT embedding cosine similarity), and information preservation (MiniCheck factual consistency). We include both small- and large-scale variants for each fusion family: early fusion (Llama 4 Scout, Llama 4 Maverick), middle fusion (Llama 3.2-Vision 11B, Llama 3.2-Vision 90B), and late fusion (Qwen2.5-VL-7B, Qwen2.5-VL-72B).

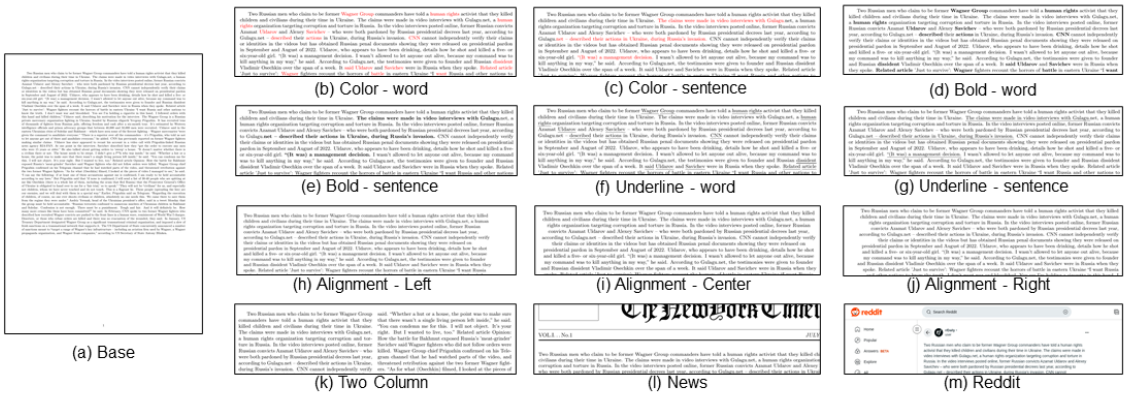

Figure 4: Examples of visual text transformations applied to the same document content. (a) Base: original document layout. (b)-(g) Text Emphasis transformations: color, bold, underline emphasis applied at word and sentence levels. (h)-(m) Structure Formatting transformations: alignment variations (left, center, right), two-column layout, and design (news article format, Reddit post format).

C Detailed Statistical Results

We further report paired significance tests comparing the base condition against each visual transformation for the same three similarity dimensions. Following standard practice, we use paired t-tests when assumptions are met (reported as normal values) and Wilcoxon signed-rank tests otherwise (reported as underlined values).

(a) Text Emphasis		Color		Bold		Underline	
		Word	Sentence	Word	Sentence	Word	Sentence
Lexical	Mean	0.60	0.59	0.59	0.59	0.58	0.59
	std	0.15	0.13	0.12	0.12	0.12	0.13
Semantic	Mean	0.77	0.77	0.78	0.78	0.77	0.77
	std	0.13	0.11	0.11	0.11	0.12	0.12
Information	Mean	0.47	0.49	0.48	0.50	0.48	0.49
	std	0.33	0.32	0.32	0.32	0.33	0.33

(b) Structure Formatting		Alignment			Column	Design	
		Left	Center	Right	Two	Reddit	News
Lexical	Mean	0.59	0.60	0.59	0.59	0.54	0.57
	std	0.13	0.14	0.16	0.15	0.15	0.17
Semantic	Mean	0.76	0.77	0.75	0.76	0.70	0.72
	std	0.12	0.13	0.15	0.14	0.17	0.17
Information	Mean	0.47	0.50	0.49	0.46	0.44	0.20
	std	0.33	0.32	0.33	0.33	0.30	0.27

Table 6: Llama 4 Scout mean similarity scores and standard deviations between original and visually transformed document summaries, measured across lexical (TF-IDF cosine similarity), semantic (Sentence-BERT embeddings), and information preservation (MiniCheck factual consistency) dimensions.

(a) Text Emphasis		Color		Bold		Underline	
		Word	Sentence	Word	Sentence	Word	Sentence
Lexical	Mean	0.50	0.52	0.54	0.52	0.52	0.52
	std	0.23	0.23	0.25	0.23	0.24	0.25
Semantic	Mean	0.65	0.66	0.66	0.66	0.65	0.64
	std	0.18	0.17	0.18	0.17	0.18	0.19
Information	Mean	0.15	0.15	0.18	0.18	0.18	0.19
	std	0.15	0.15	0.17	0.17	0.17	0.19

(b) Structure Formatting		Alignment			Column	Design	
		Left	Center	Right	Two	Reddit	News
Lexical	Mean	0.55	0.50	0.51	0.49	0.40	0.46
	std	0.25	0.25	0.24	0.23	0.27	0.24
Semantic	Mean	0.66	0.62	0.64	0.63	0.51	0.62
	std	0.18	0.20	0.19	0.20	0.23	0.20
Information	Mean	0.19	0.17	0.18	0.14	0.08	0.04
	std	0.18	0.16	0.18	0.15	0.08	0.06

Table 7: Llama 3.2-Vision 11B mean similarity scores and standard deviations between original and visually transformed document summaries, measured across lexical (TF-IDF cosine similarity), semantic (Sentence-BERT embeddings), and information preservation (MiniCheck factual consistency) dimensions.

(a) Text Emphasis		Color		Bold		Underline	
		Word	Sentence	Word	Sentence	Word	Sentence
Lexical	Mean	0.64	0.61	0.62	0.60	0.60	0.60
	std	0.15	0.13	0.13	0.12	0.13	0.12
Semantic	Mean	0.87	0.86	0.87	0.86	0.86	0.87
	std	0.08	0.09	0.06	0.07	0.09	0.07
Information	Mean	0.67	0.66	0.71	0.72	0.69	0.72
	std	0.25	0.27	0.20	0.22	0.22	0.21
(b) Structure Formatting		**Alignment**			**Column**	**Design**	
		Left	Center	Right	Two	Reddit	News
Lexical	Mean	0.62	0.63	0.62	0.62	0.59	0.58
	std	0.14	0.13	0.13	0.13	0.12	0.10
Semantic	Mean	0.87	0.87	0.87	0.87	0.86	0.86
	std	0.08	0.08	0.09	0.08	0.08	0.07
Information	Mean	0.69	0.69	0.70	0.70	0.71	0.71
	std	0.24	0.22	0.24	0.21	0.21	0.23

Table 8: Qwen2.5-VL-7B mean similarity scores and standard deviations between original and visually transformed document summaries, measured across lexical (TF-IDF cosine similarity), semantic (Sentence-BERT embeddings), and information preservation (MiniCheck factual consistency) dimensions.

(a) Text Emphasis		Color		Bold		Underline	
		Word	Sentence	Word	Sentence	Word	Sentence
Lexical	Mean	0.65	0.65	0.65	0.66	0.65	0.65
	std	0.12	0.13	0.12	0.11	0.13	0.11
Semantic	Mean	0.83	0.83	0.83	0.84	0.83	0.83
	std	0.10	0.10	0.08	0.08	0.10	0.08
Information	Mean	0.22	0.20	0.21	0.19	0.20	0.20
	std	0.27	0.23	0.25	0.23	0.23	0.23
(b) Structure Formatting		**Alignment**			**Column**	**Design**	
		Left	Center	Right	Two	Reddit	News
Lexical	Mean	0.65	0.66	0.67	0.65	0.62	0.63
	std	0.13	0.12	0.13	0.14	0.13	0.16
Semantic	Mean	0.82	0.82	0.82	0.81	0.73	0.73
	std	0.09	0.09	0.09	0.11	0.13	0.13
Information	Mean	0.21	0.22	0.22	0.20	0.20	0.23
	std	0.21	0.23	0.25	0.22	0.24	0.27

Table 9: LLaMA-4 Maverick mean similarity scores and standard deviations between original and visually transformed document summaries, measured across lexical (TF-IDF cosine similarity), semantic (Sentence-BERT embeddings), and information preservation (MiniCheck factual consistency) dimensions.

(a) Text Emphasis		Color		Bold		Underline	
		Word	Sentence	Word	Sentence	Word	Sentence
Lexical	Mean	0.67	0.67	0.67	0.68	0.66	0.66
	std	0.12	0.13	0.13	0.12	0.12	0.12
Semantic	Mean	0.83	0.82	0.82	0.83	0.82	0.83
	std	0.09	0.09	0.09	0.08	0.09	0.08
Information	Mean	0.23	0.22	0.23	0.21	0.21	0.21
	std	0.26	0.28	0.27	0.23	0.25	0.23

(b) Structure Formatting		Alignment			Column	Design	
		Left	Center	Right	Two	Reddit	News
Lexical	Mean	0.67	0.68	0.67	0.65	0.62	0.64
	std	0.11	0.12	0.13	0.13	0.14	0.15
Semantic	Mean	0.83	0.81	0.82	0.80	0.72	0.72
	std	0.08	0.10	0.09	0.10	0.12	0.11
Information	Mean	0.22	0.23	0.22	0.20	0.20	0.24
	std	0.23	0.27	0.23	0.23	0.22	0.27

Table 10: LLaMA-3.2 90B mean similarity scores and standard deviations between original and visually transformed document summaries, measured across lexical (TF-IDF cosine similarity), semantic (Sentence-BERT embeddings), and information preservation (MiniCheck factual consistency) dimensions.

(a) Text Emphasis		Color		Bold		Underline	
		Word	Sentence	Word	Sentence	Word	Sentence
Lexical	Mean	0.66	0.67	0.65	0.65	0.66	0.65
	std	0.09	0.09	0.10	0.09	0.10	0.09
Semantic	Mean	0.88	0.89	0.88	0.88	0.88	0.88
	std	0.06	0.06	0.06	0.06	0.06	0.06
Information	Mean	0.20	0.22	0.20	0.21	0.19	0.21
	std	0.21	0.23	0.21	0.20	0.20	0.21

(b) Structure Formatting		Alignment			Column	Design	
		Left	Center	Right	Two	Reddit	News
Lexical	Mean	0.66	0.66	0.66	0.65	0.64	0.63
	std	0.09	0.09	0.09	0.10	0.10	0.10
Semantic	Mean	0.89	0.89	0.88	0.87	0.87	0.88
	std	0.06	0.06	0.06	0.07	0.06	0.06
Information	Mean	0.20	0.20	0.20	0.20	0.20	0.21
	std	0.20	0.19	0.19	0.21	0.19	0.21

Table 11: Qwen2.5 72B mean similarity scores and standard deviations between original and visually transformed document summaries, measured across lexical (TF-IDF cosine similarity), semantic (Sentence-BERT embeddings), and information preservation (MiniCheck factual consistency) dimensions.

(a) Text Emphasis		Color		Bold		Underline	
		Word	Sentence	Word	Sentence	Word	Sentence
Lexical	Stats	2515	1915	1.79	1983	1916	2046
	p-value	0.97	< 0.05	0.08	0.06	< 0.05	0.10
Semantic	Stats	2171	-0.46	2123	2089	2125	2158
	p-value	0.22	0.65	0.17	0.13	0.17	0.21
Information	Stats	4284	4846	4312	4660	3967	4998
	p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

(b) Structure Formatting		Alignment			Column	Design	
		Left	Center	Right	Two	Reddit	News
Lexical	Stats	2309	0.41	2474	2466	1616	2440
	p-value	0.46	0.68	0.86	0.84	< 0.01	0.77
Semantic	Stats	2474	2434	2439	2437	1859	2218
	p-value	0.86	0.75	0.77	0.76	< 0.05	0.30
Information	Stats	4697	5648	4789	4369	3814	831
	p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Table 12: Llama 4 Scout statistical test results comparing similarity scores between base summary and transformed image summaries across lexical, semantic, and information preservation dimensions, with test statistics and p-values from paired t-tests (normal values) and Wilcoxon signed-rank tests (underlined values).

(a) Text Emphasis		Color		Bold		Underline	
		Word	Sentence	Word	Sentence	Word	Sentence
Lexical	Stats	2.22	2.55	0.38	1.50	2.76	2.72
	p-value	< 0.05	< 0.05	0.70	0.14	< 0.01	< 0.01
Semantic	Stats	2.88	1663	1949	2.54	1653	1551
	p-value	< 0.01	< 0.01	< 0.05	< 0.05	< 0.01	< 0.001
Information	Stats	2985	2808	3982	3929	4451	4984
	p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

(b) Structure Formatting		Alignment			Column	Design	
		Left	Center	Right	Two	Reddit	News
Lexical	Stats	1.67	1597	2.57	2.15	1802	2188
	p-value	0.10	< 0.01	< 0.05	< 0.05	< 0.05	0.25
Semantic	Stats	1530	1205	1578	1664	7.01	1952
	p-value	< 0.001	< 0.001	< 0.01	< 0.01	< 0.001	< 0.05
Information	Stats	4857	3958	4020	2418	734	279
	p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Table 13: Llama 3.2-Vision 11B statistical test results comparing similarity scores between base summary and transformed image summaries across lexical, semantic, and information preservation dimensions, with test statistics and p-values from paired t-tests (normal values) and Wilcoxon signed-rank tests (underlined values).

(a) Text Emphasis		Color		Bold		Underline	
		Word	Sentence	Word	Sentence	Word	Sentence
Lexical	Stats	1801	2.62	1502	4.72	5.00	1261
	p-value	< 0.05	< 0.05	< 0.001	< 0.001	< 0.001	< 0.001
Semantic	Stats	1921	2037	1.97	1691	1534	2060
	p-value	< 0.05	0.09	0.05	< 0.01	< 0.001	0.11
Information	Stats	8250	8859	8909	8868	8623	8939
	p-value	< 0.05	0.15	0.73	0.15	0.41	0.18

(b) Structure Formatting		Alignment			Column	Design	
		Left	Center	Right	Two	Reddit	News
Lexical	Stats	981	1529	1340	1297	3.43	1382
	p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Semantic	Stats	2049	2040	1702	1944	1457	2107
	p-value	0.10	0.10	< 0.01	< 0.05	< 0.001	0.15
Information	Stats	10016	9551	9441	9985	9827	9700
	p-value	0.97	0.54	0.46	0.94	0.79	0.67

Table 14: Qwen2.5-VL-7B statistical test results comparing similarity scores between base summary and transformed image summaries across lexical, semantic, and information preservation dimensions, with test statistics and p-values from paired t-tests (normal values) and Wilcoxon signed-rank tests (underlined values).

(a) Text Emphasis		Color		Bold		Underline	
		Word	Sentence	Word	Sentence	Word	Sentence
Lexical	Stats	9434	9024	9860	9638	9042	9594
	p-value	0.45	0.21	0.82	0.62	0.22	0.58
Semantic	Stats	8636	8186	9864	7759	9063	9812
	p-value	0.08	< 0.05	0.82	< 0.01	0.23	0.77
Information	Stats	9788	8942	9111	8925	8769	9558
	p-value	0.75	0.18	0.25	0.17	0.12	0.55

(b) Structure Formatting		Alignment			Column	Design	
		Left	Center	Right	Two	Reddit	News
Lexical	Stats	9704	9537	9852	9375	6979	8166
	p-value	0.67	0.53	0.81	0.41	< 0.001	< 0.05
Semantic	Stats	9818	9094	9514	7915	3119	3015
	p-value	0.78	0.24	0.51	< 0.01	< 0.001	< 0.001
Information	Stats	9058	10024	9226	7118	659	583
	p-value	0.23	0.97	0.31	< 0.001	< 0.001	< 0.001

Table 15: LLaMA-4 Maverick statistical test results comparing similarity scores between base summary and transformed image summaries across lexical, semantic, and information preservation dimensions, with test statistics and p-values from paired t-tests (normal values) and Wilcoxon signed-rank tests (underlined values).

(a) Text Emphasis		Color		Bold		Underline	
		Word	Sentence	Word	Sentence	Word	Sentence
Lexical	Stats	9402	-0.76	-0.55	8454	9165	9893
	p-value	0.43	0.45	0.58	0.05	0.28	0.85
Semantic	Stats	9602	9281	9779	8864	9881	8915
	p-value	0.58	0.35	0.74	0.15	0.84	0.20
Information	Stats	8756	8617	9570	8375	9920	9428
	p-value	0.11	0.08	0.56	< 0.05	0.87	0.45

(b) Structure Formatting		Alignment			Column	Design	
		Left	Center	Right	Two	Reddit	News
Lexical	Stats	-1.27	-0.58	-1.55	1.40	4.65	1.64
	p-value	0.21	0.56	0.12	0.16	< 0.001	0.10
Semantic	Stats	9599	8429	9540	7965	10.23	1974
	p-value	0.58	< 0.05	0.53	< 0.05	< 0.001	< 0.001
Information	Stats	9738	10018	9619	4.34	10	265
	p-value	0.70	0.97	0.60	< 0.001	< 0.001	< 0.001

Table 16: LLaMA-3.2 90B statistical test results comparing similarity scores between base summary and transformed image summaries across lexical, semantic, and information preservation dimensions, with test statistics and p-values from paired t-tests (normal values) and Wilcoxon signed-rank tests (underlined values).

(a) Text Emphasis		Color		Bold		Underline	
		Word	Sentence	Word	Sentence	Word	Sentence
Lexical	Stats	9225	8510	8883	8919	9752	8120
	p-value	0.31	0.06	0.15	0.17	0.72	< 0.05
Semantic	Stats	9596	8820	-0.35	9847	9510	9482
	p-value	0.58	0.13	0.72	0.80	0.51	0.49
Information	Stats	9496	8909	9374	9055	9563	9691
	p-value	0.50	0.16	0.48	0.22	0.82	0.66

(b) Structure Formatting		Alignment			Column	Design	
		Left	Center	Right	Two	Reddit	News
Lexical	Stats	9425	9581	9457	8498	7069	6128
	p-value	0.45	0.57	0.60	0.47	< 0.001	< 0.001
Semantic	Stats	8255	7723	9813	8534	7905	9513
	p-value	< 0.05	< 0.01	0.77	0.06	< 0.01	0.51
Information	Stats	9659	9995	9238	8933	7746	3820
	p-value	0.63	0.95	0.32	0.17	< 0.01	< 0.001

Table 17: Qwen2.5 72B statistical test results comparing similarity scores between base summary and transformed image summaries across lexical, semantic, and information preservation dimensions, with test statistics and p-values from paired t-tests (normal values) and Wilcoxon signed-rank tests (underlined values).