An Empirical Study of Position Bias in Visualized Text Retrieval

Cuiqiaoshu Pan[1] **Ziyang Zeng**[1*] **Dun Zhang**[2] **Yuqing Yang**[1†]
[1]Beijing University of Posts and Telecommunications [2]Prior Shape
{pancuiqiaoshu, ziyang1060, yangyuqing}@bupt.edu.cn
dunnzhang0@gmail.com

Abstract

Visualized text retrieval, which renders text into images and retrieves them with vision-language models, has been proposed to avoid lossy text parsing and compress long contexts. We ask whether it also bypasses a known flaw of text retrieval: *position bias*, the tendency to overweight a passage's beginning. Rendering 1D text into a 2D image may change the bias in two respects: its **shape**, which position is favored, and its **magnitude**, how strongly performance varies across positions. We test this by rendering the PosIR benchmark as controlled plain text images and comparing eight visual retrievers with four text retrievers on identical instances. We find that (1) position bias persists, with all visual retrievers exhibiting measurable bias (macro PSI 0.061–0.168); (2) its shape is transformed: seven of eight visual retrievers favor the middle of the page in an inverted-U pattern, contrasting with both the primacy bias of text retrievers and the lost-in-the-middle effect of LLMs with long contexts; (3) its magnitude is sensitive to rendering settings, as taller images amplify PSI and can even reverse positional preference, while font size has a similarly large effect; and (4) under matched backbones, visual retrievers exhibit lower position bias than their text counterparts on longer documents, but this advantage vanishes beyond the pixel budget and at document scale. Visualized text retrieval therefore does not eliminate position bias; it transforms a familiar 1D textual bias into a more complex 2D spatial bias.

1 Introduction

Long-context processing remains constrained by token efficiency, motivating a growing line of work that represents text through the *visual* modality. Instead of processing text directly as token sequences, these approaches render text into images and process the resulting visual inputs with vision-language models (VLMs). Optical context compression has reported substantial reductions in token usage (Wei et al., 2025; Cheng et al., 2025; Xing et al., 2026; Wang et al., 2026; Li et al., 2025; Zhong et al., 2026), and the same idea extends to *visual document retrieval*, where documents are embedded from page images (Faysse et al., 2025; Yu et al., 2025; Ma et al., 2024). Visualized text processing is therefore motivated by two potential benefits: avoiding parsing loss and compressing long contexts.

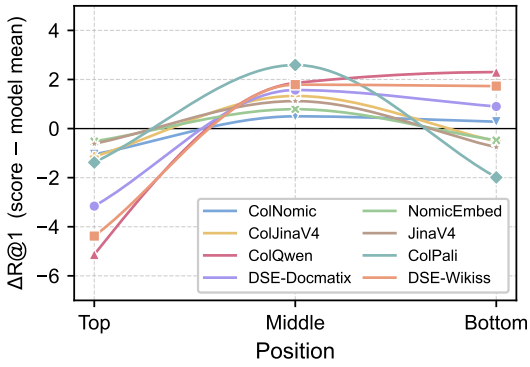


Figure 1: Overview of the positional preference of visual text retrievers on rendered PosIR. Each line shows one model's Recall@1 deviation from its own overall mean across three vertical positions, where positive values indicate above-average performance for that model. Most visual retrievers peak at the middle position, suggesting that rendering text into images changes the dominant retrieval bias from the top-favoring pattern commonly observed in text retrieval to a middle-favoring spatial pattern. Detailed per-position scores and PSI values are reported in Table 1.

We ask whether this visual representation also bypasses a known weakness of text retrieval: *position bias*. Prior work has shown that dense and late-interaction text retrievers tend to overweight evidence near the beginning of a passage and underweight equally relevant evidence appearing later

[*] Project lead.
[†] Corresponding author.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5740–5760
November 6-10, 2026 ©2026 Association for Computational Linguistics

(Zeng et al., 2025, 2026; Coelho et al., 2024; Lee et al., 2025). When text is rendered as a fixed-width, variable-height image, its 1D token sequence becomes a 2D spatial layout. This transformation may alter both the **shape** of position bias, which positions are favored, and its **magnitude**, how strongly retrieval performance varies across positions. Yet how position bias changes under this visual representation remains largely unexplored.

We investigate this question empirically using PosIR (Zeng et al., 2026), a benchmark designed to measure position bias. We render its documents into controlled plain text images and evaluate eight visual retrievers against four text retrievers on identical instances and shared position buckets. For controlled cross-modal comparison, we additionally include two pairs of matched-backbone visual and text retrievers from the Jina and Nomic model families. As a premise, position bias persists after visualization: all eight visual retrievers exhibit non-trivial bias, with macro PSI ranging from 0.061 to 0.168. We further conduct controlled experiments and saliency analyses to examine how rendering shapes position bias, explore its underlying mechanisms, and characterize the performance boundaries of the visual retrieval advantage. Our main findings are:

- **F1: The shape of position bias is transformed.** Visual retrievers generally exhibit a middle preference, forming a novel inverted-U pattern distinct from previously observed biases. On identical instances, seven of eight visual retrievers score highest at the Middle position, while text retrievers perform best at the Top position.
- **F2: The magnitude of position bias is rendering-dependent.** Increasing image height amplifies position bias and can reverse positional preference. Rendering parameters also affect model behavior, with font size showing the strongest effect.
- **F3: Visualized text retrieval is more position-stable.** Across the two matched-backbone pairs, visual models degrade more gradually and exhibit lower position bias than text counterparts. This advantage becomes more pronounced as document length increases.
- **F4: This visual advantage is bounded.** It does not hold across all input lengths and retrieval tasks. Results on LongEmbed and BEIR suggest that the visual input pixel budget and differences in lexical-overlap sensitivity may contribute to these limitations.

Taken together, these findings point to a new positional failure mode in visualized text retrieval: the relevant page may be retrieved, yet evidence at particular locations within it is systematically underweighted.

2 Related Work

Visual and optical compression. Rendering text into images and reading it back with VLMs has emerged as a way to compress long context: DeepSeek-OCR (Wei et al., 2025) and Glyph (Cheng et al., 2025) report large token-compression ratios, and the idea has been extended to in-context learning, reasoning, and general inputs (Xing et al., 2026; Wang et al., 2026; Li et al., 2025). A parallel critical line questions how much of the reported quality is genuine, arguing that downstream scores are inflated by language priors or that optical compression is a poor autoencoder (Liu and Qiu, 2025; Lee et al., 2026; Zhao et al., 2025; Gao et al., 2026; Liang et al., 2026). These audits target *decoding* fidelity; we audit *retrieval-side positional* fidelity.

Visual document retrieval. ColPali (Faysse et al., 2025) embeds document page images directly with a VLM and outperforms text pipelines on visually rich documents; VisRAG (Yu et al., 2025) and document screenshot embedding (Ma et al., 2024) build dense vision-based retrieval pipelines. VISTA-Bench (Liu et al., 2026) provides a controlled rendering protocol for visualized text, closely related to our setup. These models are our objects of study; existing evaluations report aggregate accuracy but not how it varies with the position of relevant content.

Position bias in retrieval. Zeng et al. (2025) formalize position bias across the retrieval pipeline and propose the Position Sensitivity Index (PSI); PosIR (Zeng et al., 2026) extends this to a heterogeneous, length-controlled benchmark. Primacy bias in text embedding models is documented by Coelho et al. (2024) and Lee et al. (2025), while "lost-in-the-middle" (Liu et al., 2024; Hsieh et al., 2024) describes a U-shaped bias in long-context LLMs. Closest to us, Wu and Albreiki (2025) study position bias in multimodal embedding models, but in an inverted setting: they perturb text position in the

query against natural images, using single-vector contrastive models not optimized for retrieval. We instead render the corpus into text images, keep the query textual, evaluate actual visual document retrievers including multi-vector late-interaction models, and preserve document reading order—porting the PSI framework to a 2D spatial layout.

3 Methodology

Task. Our object of study is visualized text retrieval: a textual query q is matched against a collection of document images $\mathcal{D} = \{d_1, \ldots, d_N\}$, each encoded by a VLM and ranked by a relevance score $s(q, d)$. We focus on page-level retrieval—identifying which image contains the relevant information. For the text retrievers used as a reference, the query is identical; only the corpus differs, consisting of the original source documents rather than their rendered images.

Data and controlled rendering. We build on PosIR (Zeng et al., 2026), a retrieval benchmark purpose-built to study position bias: it spans 31 domains, each instance pairs a document and query with a reference span, documents preserve their original reading order, and reference spans are evenly distributed across positions by design. To isolate position from content and layout confounds, we render each document into a plain-text image with a uniform pipeline (PIL; width fixed at 1000 px, height grows with content; Arial 24 px; $1.5\times$ line spacing) stripped of headers, figures, tables, columns, and page numbers. To probe how typography interacts with position bias, we additionally vary one parameter at a time: font size (12–48 px), line spacing (1.0–$2.0\times$), and typeface (5 fonts). To make the rendering pipeline concrete, Figure 2 shows a single PosIR passage rendered under the baseline configuration and two single-parameter variants. Beyond PosIR, we use BEIR (Thakur et al., 2021) and LongEmbed (Zhu et al., 2024) to delimit the scope of our findings (§4.4), both rendered through the same pipeline.

Position annotation. For each rendered image we record the bounding box of the relevant region, compute the vertical center of this bounding box, normalize it to $[0, 1]$ by image height, and assign it to **Top** $[0, 0.33)$, **Middle** $[0.33, 0.67)$, or **Bottom** $[0.67, 1.0]$. A relevant region that spans a bucket boundary is assigned by the location of its center, not its extent.

Modality-controlled bucketing. Visual and text retrievers are scored against the same per-instance position bucket, defined by the y-coordinate of the relevant region in the rendered image. Because rendering is linear top-to-bottom with uniform typography, this coordinate tracks the textual position of the relevant span. Therefore, when visual and text retrievers show different PSI values, the difference comes from how the two retrieval modalities respond to the same position split, not from using different definitions of position.

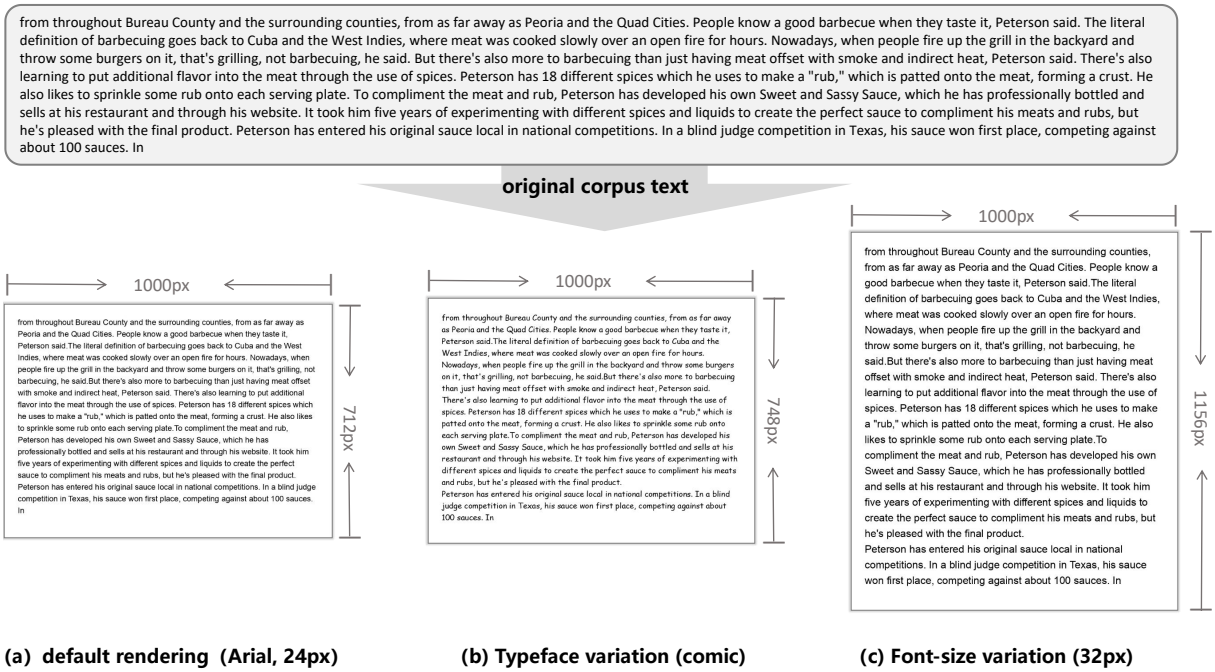
Metrics. We report Recall@1 (R@1) and NDCG@5, following standard practice in visual document retrieval (Faysse et al., 2025), and quantify position bias with the Position Sensitivity Index (PSI) (Zeng et al., 2025):

$$\mathrm{PSI} = \frac{S_{\max} - S_{\min}}{S_{\max}}, \quad S \in \{S_{\text{Top}}, S_{\text{Mid}}, S_{\text{Bot}}\} \tag{1}$$

where $S_{\max}$ and $S_{\min}$ are the highest and lowest scores across the three position buckets. $\mathrm{PSI} = 0$ indicates position-invariant performance; larger values indicate stronger bias. We report micro PSI (globally pooled buckets) and macro PSI (mean of per-domain PSI), with macro PSI as primary, and additionally report the weakest position as an indicator of bias direction. Because PSI is normalized by $S_{\max}$, it is coupled with overall accuracy: a more accurate model may obtain a smaller PSI partly because its best-position score is larger. Therefore, within-model comparisons across image height or rendering settings are cleaner than rankings of PSI magnitude. We treat cross-model PSI comparisons as descriptive and interpret them together with overall retrieval accuracy.

Models. We evaluate eight visual document retrievers grouped by output mode: **late-interaction** (ColPali and ColQwen (Faysse et al., 2025), ColJinaV4 (Günther et al., 2025), ColNomic (Nomic AI, 2025)) and **dense**(DSE-Docmatix and DSE-Wikiss (Ma et al., 2024), JinaV4 (Günther et al., 2025), NomicEmbed (Nussbaum et al., 2024)).

We also evaluate four text retrievers as a 1D reference: BGE-M3 (Chen et al., 2025), BGE-Reranker-v2-m3 (Chen et al., 2025; BAAI, 2024), JinaV4-text (Günther et al., 2025), and NomicEmbed-text (Nussbaum et al., 2025). The last two share backbones with their visual counterparts. All text retrievers are run on the PosIR source text and are evaluated using the same position buckets as the visual retrievers.



(a) default rendering (Arial, 24px) **(b) Typeface variation (comic)** **(c) Font-size variation (32px)**

Figure 2: An illustration of the rendering pipeline. A single source passage from PosIR (top) is rendered into a plain-text image under three configurations: (a) the baseline—Arial, 24 px, 1.5× line spacing; (b) a typeface variation—the same text set in a handwriting-style (Comic) font; and (c) a font-size variation—Arial enlarged to 32 px. All three keep the canvas width fixed at 1000 px while the height grows with content (712, 748, and 1156 px respectively), so that a larger font or a less compact typeface yields a taller image of the same text. Headers, figures, tables, columns, and page numbers are stripped, leaving only the linearly laid-out body text.

Protocols. The full-scale protocol is our primary evaluation: all 31 domains under baseline rendering, macro-averaged across domains (mean ± standard deviation). The sampled protocol, used for the rendering experiments in §4.3.2, draws 300 stratified samples per position bucket from 10 randomly selected domains, averaged over 5 random seeds.

4 Findings

We organize our findings as a three-act account of how visualization reshapes position bias. We first establish a premise: position bias does not vanish under visualization (§4.1). We then characterize the bias along two axes—its **shape**, namely which position is favored (§4.2), and its **magnitude**, namely how large the gap across positions is and what drives it (§4.3). Finally, we delimit the **boundary** within which visualized retrieval is the more position-stable modality, and show where that advantage breaks down (§4.4).

4.1 Position Bias Persists in Visualized Text Retrieval

Table 1 reports full-scale results for both modalities on the same instances and the same per-instance position buckets. All eight visual retrievers exhibit non-trivial position bias: macro PSI on Recall@1 ranges from 0.061 (ColNomic) to 0.168 (DSE-Wikiss). In other words, rendering text into images does not make retrieval position-invariant. A randomization test within domains further supports position dependence in all eight visual retrievers, with Holm-adjusted p-values ≤ 0.0014 (Appendix B.3). As a 1D reference, the four text retrievers all show a primacy bias, consistent with prior work on text retrieval (Zeng et al., 2025; Coelho et al., 2024); their macro-PSI values are higher than those of most visual retrievers. Visualizing text thus does not remove position bias, though most visual retrievers have lower PSI than the text baselines on the same instances.

4.2 Shape: Visualized Retrieval Exhibits a Middle Preference

We begin with the shape of the bias. Text retrievers inherit a primacy bias: on the text retrievers in Table 1, Top is the strongest position on the very same instances. For seven out of the eight visual retrievers we tested, the Middle position yields the highest score, outperforming the better result of the Top and Bottom positions by margins ranging from +0.06 (DSE-Wikiss) to +3.97 (ColPali). ColJinaV4,

Modality	Model	Overall		Per-position			Bias Metrics		
		R@1	NDCG@5	Top	Middle	Bottom	Δ_{Middle}	PSI (R)	PSI (N)
Text	BGE-Reranker-v2	64.39 ± 5.47	73.11 ± 4.27	79.36	63.89	51.38	-15.47	0.352	0.310
	JinaV4-text	48.04 ± 7.06	59.65 ± 6.63	51.67	47.80	45.06	-3.87	0.128	0.119
	NomicEmbed-text	43.87 ± 6.23	54.46 ± 6.13	49.65	43.74	38.79	-5.91	0.219	0.192
	BGE-M3	49.59 ± 7.05	61.24 ± 6.45	53.91	48.87	46.56	-5.04	0.154	0.126
Visual	ColNomic	77.59 ± 5.61	87.35 ± 3.50	76.54	78.09	77.87	$+0.22$	0.061	0.033
	ColJinaV4	74.12 ± 6.31	84.24 ± 4.67	72.98	75.45	73.60	$+1.85$	0.064	0.037
	ColQwen	65.16 ± 5.43	76.00 ± 4.23	60.03	67.02	67.46	-0.44	0.125	0.093
	DSE-Docmatix	55.73 ± 7.36	67.28 ± 6.72	52.57	57.30	56.63	$+0.67$	0.115	0.084
	NomicEmbed	55.60 ± 5.65	67.52 ± 4.68	55.08	56.39	55.12	$+1.27$	0.091	0.057
	JinaV4	53.74 ± 7.30	65.64 ± 6.43	53.12	54.86	52.98	$+1.74$	0.100	0.075
	ColPali	53.51 ± 6.05	63.38 ± 5.23	52.13	56.10	51.52	$+3.97$	0.121	0.106
	DSE-Wikiss	42.52 ± 5.85	53.93 ± 5.80	38.14	44.31	44.25	$+0.06$	0.168	0.128

Table 1: Full-scale results on rendered PosIR. Visual retrievers are sorted by overall Recall@1; text retrievers appear above the rule as a 1D reference. Both modalities are evaluated on identical instances and identical per-instance position buckets (defined by the y-coordinate of the relevant region in the rendered image; see §3); differences in PSI therefore reflect modality-level processing rather than partition definitions. Per-position columns report Recall@1. The Δ_{Middle} column reports Middle minus the higher-scoring of Top and Bottom; positive Δ_{Middle} indicates a middle-favoring (inverted-U) shape, negative indicates an end-favoring (primacy or recency) shape.

JinaV4, and ColPali show a statistically significant middle advantage, while the remaining four demonstrate a weaker middle-leaning trend with higher uncertainty at the individual-model level (Appendix B.3). Only ColQwen follows a monotonic pattern ($\Delta_{\text{Middle}} = -0.44$), with performance increasing from Top to Bottom. The key change introduced by visualization is therefore not merely a change in bias magnitude, but a change in bias direction: most visual models lean toward the middle of the rendered page rather than the beginning of the text.

This middle preference is unexpected in two respects. First, it contrasts with the primacy bias observed in text retrievers on the same instances and documented in prior work. Second, it runs counter to the lost-in-the-middle effect in long-context LLMs, where information in the middle is harder to use (Liu et al., 2024; Hsieh et al., 2024).

Before interpreting the inverted-U shape, we examine two simpler explanations. First, it is unlikely to originate from the data itself. This is because PosIR deliberately distributes relevant spans across positions and controls for retrieval difficulty, and the text retrievers evaluated on the same set of instances do not exhibit a middle advantage. In addition, Top is not consistently the weakest position across models in Figure 1, arguing against a uniformly defective Top bucket. Second, it is unlikely to be a simple rendering artifact. Our uniform rendering also removes headers, figures, tables, columns, and page numbers. As illustrated in Figure 2, different vertical regions are visually homogeneous except for variation in their positions.

To complement these controls, we conduct a position-shift experiment with 1,710 queries from four domains, moving the same relevant span to Top, Middle, and Bottom positions while keeping the query, the rest of the document text, and candidate-pool membership unchanged. We evaluate two models: ColPali, which has the largest Δ_{Middle} value in Table 1, and ColQwen, the only visual model whose aggregate score peaks at Bottom. Across both R@1 and NDCG@5, ColPali performs significantly worse at the Bottom than at the Middle, with an R@1 gap of 8.97 percentage points (95% CI: [7.02, 10.91]), while ColQwen shows smaller positional differences whose 95% confidence intervals include zero (Appendix B.4). These results support model-dependent positional sensitivity, although moving the span may also change its local discourse context. These controls suggest that the inverted-U pattern reflects the retriever rather than the dataset or layout.

4.3 Magnitude: How Large the Bias Is, and What Affects It

This section studies the magnitude of position bias and identifies two factors that shape it: the height of the rendered image (§4.3.1) and the rendering parameters of the text-to-image rendering (§4.3.2). Image height is our main focus, as it connects di-

Modality	Model	Length / Height	Recall@1				PSI
			Overall	Top	Middle	Bottom	
Text	BGE-M3	short ($\leq$1024)	71.75 ± 7.71	73.23	72.15	69.87	0.109
		medium	65.49 ± 7.28	72.64	63.51	60.33	0.178
		tall	57.05 ± 7.88	63.42	55.58	52.15	0.213
		very tall ($>$3072)	52.62 ± 6.78	57.56	49.13	51.18	0.174
	NomicEmbed-text	short	70.22 ± 8.31	71.65	70.88	68.14	0.117
		medium	62.65 ± 6.62	69.25	61.69	57.00	0.187
		tall	49.07 ± 7.08	58.57	48.00	40.63	0.315
		very tall	38.81 ± 6.60	49.02	34.67	32.76	0.361
Visual	ColPali	short ($\leq$1024)	80.66 ± 7.27	77.74	81.67	82.57	0.098
		medium	79.89 ± 5.20	79.17	80.18	80.32	0.059
		tall	62.42 ± 6.03	67.87	66.38	53.00	0.240
		very tall ($>$3072)	34.11 ± 5.63	39.57	33.90	28.85	0.296
	ColNomic	short	90.50 ± 3.34	89.51	90.84	91.17	0.049
		medium	88.96 ± 3.81	88.36	89.20	89.32	0.051
		tall	86.70 ± 3.75	86.19	86.54	87.37	0.061
		very tall	83.29 ± 4.95	84.09	83.63	82.15	0.061
	ColQwen	short	87.31 ± 4.25	84.81	88.16	88.95	0.078
		medium	86.08 ± 4.35	84.55	86.62	87.08	0.054
		tall	79.31 ± 5.21	76.80	80.08	81.06	0.083
		very tall	54.45 ± 7.63	51.11	54.42	57.83	0.156

Table 2: Recall@1 and macro PSI as a function of input length. Both modalities use one common per-instance position partition, defined on the rendered image and shared with the text side (§3); buckets are ordered by rendered image height, which under linear rendering coincides with source length. The DSE models follow patterns similar to ColQwen; the remaining text retrievers follow patterns similar to BGE-M3. Complete results are in Appendix C.

rectly to the pixel-budget limit examined in §4.4; the rendering parameters serve a secondary, supporting role.

4.3.1 Image Height Amplifies the Bias

PosIR documents vary substantially in length. Under fixed-width rendering, longer documents become taller images. This gives us a natural way to study how input length interacts with position bias in visualized retrieval.

Accuracy and bias both worsen with height. All eight visual models lose accuracy as image height grows; the Recall@1 drop from the short to very-tall bucket ranges from 7.2 points (ColNomic) to 46.6 points (ColPali). Position bias also amplifies with height, but non-uniformly across models: ColPali's macro PSI rises from 0.098 to 0.296 (a roughly 3$\times$ increase), whereas ColNomic and ColJinaV4 increase by only about +0.01–0.02 PSI between the short and very-tall buckets.

Image height also alters the model's preferred position. ColPali demonstrates the clearest reversal: the Bottom position yields the highest aggregate score for short images, while the Top position performs best for very tall images. Across all domains, the Top position is the optimal position in 4 cases within the short-image bucket and 24 cases within the very-tall bucket (Table 11). Other models, such as ColNomic, also shift toward a preference for the Top position, but this change is less pronounced. Neither the DSE models nor ColQwen exhibit a comparable reversal. By contrast, the text baselines listed in Table 2 show a preference for the Top position across every height bucket. Image height can therefore alter not only the magnitude of position bias but also the preferred position, and this height-driven position reversal is a unique characteristic of visualized text retrieval.

Within a shared backbone, the visual side is the more stable one. As image height increases, the visual models evaluated in our comparisons tend to be more stable. To reduce confounding from differences in backbone, training data and scoring mechanisms, we focus on two model pairs from the Nomic and Jina families, with each pair sharing a backbone. We investigate how the performance of these two pairs changes as the rendered image height varies. When moving from the short to the very-tall height bucket, ColNomic sees a Recall@1 drop of 7.2 points, compared to a 31.4-point drop for NomicEmbed-text. ColNomic's macro PSI stays around 0.05–0.06, while NomicEmbed-

text's macro PSI climbs from 0.117 to 0.361. The Jina comparison follows the same trend: ColJinaV4 loses 12.9 points, versus a 23.7-point loss for JinaV4-text (Appendix C). For long documents, visual models show higher accuracy and lower position bias than their text counterparts. That said, this pattern only applies to the two comparison pairs we tested and does not hold for all evaluated visual and text retrievers. As demonstrated in §4.4, this advantage also depends on evaluation conditions, and it does not establish that visualization has completely eliminated position bias.

4.3.2 Rendering Parameters Also Modulate Bias Magnitude

Image height is not the only factor that affects bias magnitude. In visualized retrieval, text-to-image compression is implemented by rendering source text onto a blank canvas, and the rendering choices made in this step—font size, line spacing, and typeface—may also shape the resulting bias. We therefore vary these parameters under the sampled protocol described in Section 3. Figure 3 reports the font-size sweep.

Font size has the strongest effect. As font size increases from 12 px to 48 px, most models in this sweep exhibit both decreased retrieval accuracy and an increased PSI. For example, ColPali's Recall@1 drops from 88.06 to 45.40, while its PSI rises from 0.076 to 0.154. This sensitivity confirms that the observed position bias is not an artifact of any single specific rendering configuration; instead, it changes systematically when adjusting the text-to-image rendering process. Line spacing and typeface have weaker and less consistent effects, and we report their full results in Appendix E.

Under fixed-width rendering, increasing the font size also increases rendered image height when the complete source text is retained. In our main experiment on font size, we kept the text unchanged across renderings. We also conducted a supplementary experiment to compare font sizes within the same height bucket (full results in Appendix E.3). In this supplementary experiment, accuracy generally improves when font size increases from very small to moderate; in some cases, further font enlargement can reduce accuracy or increase PSI. Although this comparison uses images with similar heights, those images do not necessarily come from the same document: increasing the font size may move a document into a taller height bucket.

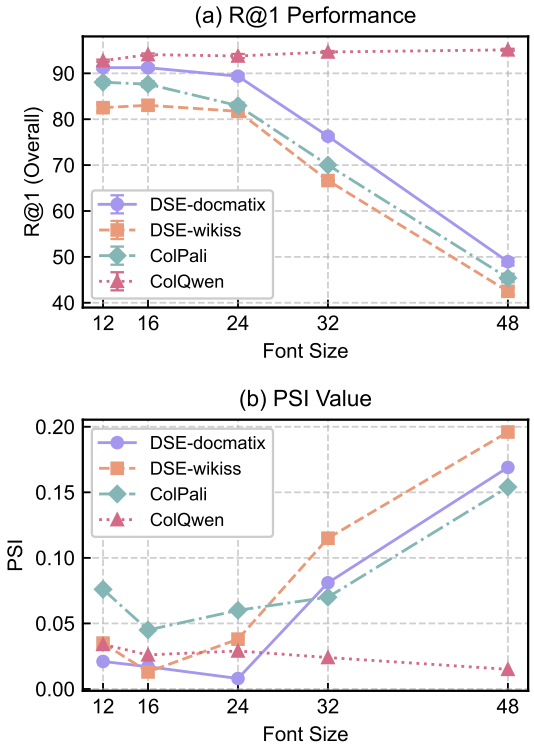


Figure 3: Font Size Effects. (a) Recall@1 as a function of font size. (b) PSI as a function of font size.

This supplementary experiment thus provides an additional comparison under similar image heights, which complements the main experiment.

Resolution strategy. The contrast between dynamic- and fixed-resolution models is also notable. For example, ColQwen (dynamic resolution) responds differently to font enlargement than ColPali (fixed resolution): ColQwen remains stable or improves as font size increases, whereas ColPali loses accuracy. This difference motivates the gradient-based analysis in §4.3.3.

4.3.3 Gradient Saliency Analysis

To better understand the mechanisms underlying these findings, we employ gradient-based saliency analysis to investigate how different text regions contribute to the retrieval score across positions and rendering settings. We focus on three questions: (1) how increasing image height strengthens position bias and reverses positional preference; (2) how font size affects position bias; and (3) how fixed- and dynamic-resolution models respond differently to changes in rendering parameters.

We conduct this analysis on ColPali and ColQwen. For each model, we calculate the gradients of the MaxSim retrieval score with respect

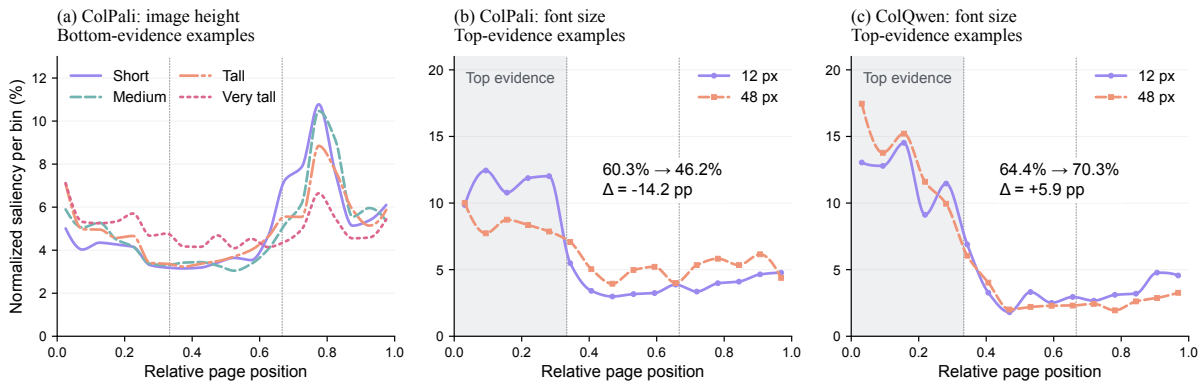


Figure 4: Effects of image height and font size on gradient-based saliency. (a) ColPali across height buckets, with relevant regions assigned to Bottom. (b–c) ColPali and ColQwen at 12 and 48 px, using the same pairs with relevant regions assigned to Top. Curves show mean normalized saliency per vertical bin: 20 bins in (a) and 16 in (b–c). Shading marks the Top position; annotations show its saliency share and change in percentage points. Gradients are computed at the first decoder-block input in (a) and the final decoder-block input in (b–c).

to the hidden states of visual tokens, extract the L_2 norm of the gradient for each token, and normalize the obtained values across all visual tokens in each rendered image (Appendix F). Figure 4 visualizes the resulting saliency distributions along the vertical axis of the rendered images, where each curve represents the average proportion of normalized saliency in each vertical bin.

To examine the positional preference reversal reported in §4.3.1, we compare ColPali saliency distributions across height buckets using samples whose relevant regions are assigned to Bottom (Figure 4(a)). For short images, the curve has a distinct peak in the Bottom region, where the evidence is located. As image height increases, this peak gradually weakens; in the very-tall bucket, the peak becomes far less pronounced, and the Top position accounts for a larger proportion of total saliency. These results are consistent with the shift from Bottom to Top preference observed for ColPali.

Figures 4(b) and 4(c) compare different font sizes using the same sample pairs with evidence in the Top region. For ColPali, increasing the font size from 12 px to 48 px weakens the saliency peak in the Top region, while saliency increases across most of the remaining area of the image (Figure 4(b)). This indicates that an increased font size makes it more difficult for the model (ColPali) to maintain focus on the relevant region containing evidence. However, as a dynamic-resolution model, ColQwen shows the opposite response on the same set of samples. Larger fonts strengthen saliency near the top of the image and increase the concentration of saliency in the relevant region (Figure 4(c)), which means ColQwen adopts a completely different processing approach when rendering parameters such as font size change. Increasing font size degrades the performance of fixed-resolution models, but helps dynamic-resolution models better locate the relevant region. These patterns align with the retrieval trends reported in §4.3.2 and may reflect differences in how the models handle resolution, though the backbone architectures and training settings also differ between the two models.

4.4 Boundaries of the Visual Advantage

The comparison between text and visual retrievers in §4.3 may suggest that visualization substantially mitigates position bias. We caution against drawing this conclusion directly. Instead, we probe two limits of the observed visual advantage: its degradation as context length increases and its limited transfer to document-level retrieval. These evaluations characterize retrieval performance rather than directly quantify position bias.

Pixel-budget boundary. The visual advantage gradually diminishes as the rendered image size increases; in this section, we examine the limit of this effect. We evaluate this boundary on the Needle and Passkey subsets of LongEmbed (Table 3). Visual encoders remain competitive in shorter contexts, but their performance drops once documents reach several thousand tokens, whereas text encoders degrade much more gradually. Under fixed-width rendering, long documents produce excessively tall images that exceed the pixel budget of VLM encoders, making it difficult to preserve

Domain	Model	Context length (tokens)			
		256	1024	4096	16384
Needle	JinaV4	87.1	25.1	4.4	0.0
	JinaV4-text	85.0	21.1	5.3	0.4
	NomicEmbed	95.3	47.0	9.1	0.0
	NomicEmbed-text	85.7	4.5	0.0	0.0
Passkey	JinaV4	100.0	45.0	4.6	0.0
	JinaV4-text	89.1	47.3	20.7	4.1
	NomicEmbed	61.4	62.7	14.1	0.0
	NomicEmbed-text	90.6	34.0	1.5	0.6

Table 3: NDCG@10 on the LongEmbed Needle and Passkey datasets as a function of context length. Visual encoders (JinaV4, NomicEmbed) are competitive at short contexts but collapse toward zero beyond a few thousand tokens. Full results across context lengths are in the appendix.

Lexical-overlap	Queries	Nomic gap	Jina gap
(a) ArguAna			
Q1: lowest	349	+5.28	+2.52
Q2	352	+5.24	+1.89
Q3	350	+6.12	−1.33
Q4: highest	350	+3.89	−0.28
(b) Pooled BEIR			
[0, 0.20]	425	+3.39	+0.88
(0.20, 0.40]	1,451	+3.30	−0.98
(0.40, 0.60]	894	−3.65	−4.93
(0.60, 0.80]	603	−2.30	−5.88
(0.80, 1.00]	299	−7.23	−5.51

Table 4: Mean per-query NDCG@10 differences between visual and text models across lexical-overlap groups, in percentage points.

text at a readable resolution. For example, a 10k-token document contains roughly 40,000 characters. With our 24 px font, it renders to an image approximately 20,000–25,000 px tall, exceeding 20 million pixels. The default JinaV4 visual input limit of 602,112 pixels requires roughly $6\times$ downsampling along each dimension, reducing 24 px text to about 4 px and making fine textual details difficult to preserve. The visual advantage is therefore constrained by the available pixel budget and disappears once the rendered input exceeds the effective resolution capacity of the visual retriever.

Task boundary. We also evaluate the same matched pairs on five BEIR datasets. In this setting, however, visual models no longer outperform their text counterparts. Text models achieve higher mean NDCG@10, and visual models outperform their text counterparts only on ArguAna (Appendix D). Our analyses suggest that lexical overlap may contribute to this variation: ArguAna has lower overlap between queries and relevant documents than the other datasets. For ArguAna and the pooled BEIR datasets, we compare the difference between visual and text retrieval scores within bins divided by lexical-overlap ranges. Both analyses exhibit a similar trend: visual models are more competitive at low lexical overlap, whereas text models grow progressively stronger as lexical overlap increases (Table 4). This suggests that direct lexical correspondence may partly account for the performance reversal observed on BEIR. Since lexical overlap may also correlate with other dataset properties, we view this result as suggestive rather than causal evidence and leave a more controlled analysis for future work. We further examine several alternative factors, including query length, document length, the number of relevant documents, and span dilution. As detailed in Appendix D, our analyses do not support these factors as explanations for the observed boundary of the visualized text retrieval advantage. Specifically, none of them explains when or why the visual advantage diminishes or disappears. Together, the two boundaries show that the visual advantage observed on PosIR does not imply a general retrieval advantage across input lengths and tasks.

5 Conclusion

Visualized text retrieval is promoted as a way to fit long text into fewer tokens by rendering it into images. Our study shows that this representation does not escape position bias; instead, it inherits the bias in a transformed form. The primacy bias of text retrieval becomes a middle-favoring spatial bias over the rendered page, and its magnitude grows with image height. Under matched backbones, the visual side carries lower position bias than the text side, but this advantage holds only within the encoder's pixel budget and does not extend to document-level retrieval. These findings carry a practical implication: a visualized-retrieval pipeline may surface the right page while systematically under-weighting evidence at particular locations on it. They also leave two questions open for future work, namely the mechanism behind the middle preference and how this retrieval-stage bias propagates into downstream generation.

Limitations

We highlight four scope limits that are directly tied to our main claims.

English subset only. Although PosIR contains multilingual data, our experiments use only its English subset. Therefore, our findings should not be taken as evidence that the same positional pattern holds across languages. This is especially relevant for languages with different scripts or reading directions, where rendering may induce different spatial layouts.

Controlled rendering rather than real document layout. Our experiments use plain-text renderings with fixed typography and without figures, tables, headers, footers, multi-column structure, or page numbers. This design is intentional: it isolates positional effects from layout and content confounds. However, it also means that our results characterize position bias under controlled visualized-text rendering, not all visually rich documents encountered in real applications.

Mechanistic explanations are correlational. We observe a middle-favoring shape in most visual retrievers and discuss possible contributing factors such as visual pretraining priors, image resizing, resolution handling, and matching granularity. These explanations remain hypotheses. Our model set also provides only limited controlled comparisons: for example, the late-interaction versus dense comparison is available for two same-backbone pairs, and resolution strategy is not isolated independently of backbone and training data. A causal analysis would require controlled model variants that differ in only one architectural factor.

Retrieval-stage scope. Our experiments evaluate the retrieval stage directly, using Recall@1, NDCG@5, and PSI. They do not measure how the retrieved visualized text is subsequently used by a generator or reader model. Thus, our conclusions are about position bias in visualized text retrieval itself, rather than the full behavior of an end-to-end visual text-compression pipeline.

References

BAAI. 2024. BGE-Reranker-v2-M3. https://huggingface.co/BAAI/bge-reranker-v2-m3.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2025. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jiale Cheng, Yusen Liu, Xinyu Zhang, Yulin Fei, Wenyi Hong, Ruiliang Lyu, Weihan Wang, Zhe Su, Xiaotao Gu, Xiao Liu, Yushi Bai, Jie Tang, Hongning Wang, and Minlie Huang. 2025. Glyph: Scaling context windows via visual-text compression. *Preprint*, arXiv:2510.17800.

João Coelho, Bruno Martins, Joao Magalhaes, Jamie Callan, and Chenyan Xiong. 2024. Dwell in the beginning: How language models embed long documents for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–377, Bangkok, Thailand. Association for Computational Linguistics.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pages 61424–61449.

Yonghan Gao, Zehong Chen, Lijian Xu, Jingzhi Chen, Jingwei Guan, and Xingyu Zeng. 2026. Zerosense:how vision matters in long context compression. *Preprint*, arXiv:2603.11846.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. *Preprint*, arXiv:2506.18902.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995, Bangkok, Thailand. Association for Computational Linguistics.

Ivan Yee Lee, Cheng Yang, and Taylor Berg-Kirkpatrick. 2026. Optical context compression is just (bad) autoencoding. *Preprint*, arXiv:2512.03643.

Reagan J. Lee, Samarth Goel, and Kannan Ramchandran. 2025. Quantifying positional biases in text embedding models. *Preprint*, arXiv:2412.15241.

Yanhong Li, Zixuan Lan, and Jiawei Zhou. 2025. Text or pixels? it takes half: On the token efficiency of visual text inputs in multimodal llms. *Preprint*, arXiv:2510.18279.

Yunhao Liang, Ruixuan Ying, Bo Li, Hong Li, Kai Yan, Qingwen Li, Min Yang, Okamoto Satoshi, Zhe Cui, and Shiwen Ni. 2026. Visual merit or linguistic crutch? a close look at deepseek-ocr. *Preprint*, arXiv:2601.03714.

Fanfan Liu and Haibo Qiu. 2025. Context cascade compression: Exploring the upper limits of text compression. *Preprint*, arXiv:2511.15244.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Qing'an Liu, Juntong Feng, Yuhao Wang, Xinzhe Han, Yujie Cheng, Yue Zhu, Haiwen Diao, Yunzhi Zhuge, and Huchuan Lu. 2026. Vista-bench: Do vision-language models really understand visualized text as well as pure text? *Preprint*, arXiv:2602.04802.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, Miami, Florida, USA. Association for Computational Linguistics.

Nomic AI. 2025. Nomic embed multimodal: Open source multimodal embedding models for text, images, pdfs, and charts. https://www.nomic.ai/news/nomic-embed-multimodal.

Zach Nussbaum, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed vision: Expanding the latent space. *Preprint*, arXiv:2406.18587.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2025. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Yibo Wang, Yongcheng Jing, Shunyu Liu, Hao Guan, Rong cheng Tu, Chengyu Wang, Jun Huang, and Dacheng Tao. 2026. Vtc-r1: Vision-text compression for efficient long-context reasoning. *Preprint*, arXiv:2601.22069.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *Preprint*, arXiv:2510.18234.

Kebin Wu and Fatima Albreiki. 2025. Positional bias in multimodal embedding models: Do they favor the beginning, the middle, or the end? *Preprint*, arXiv:2511.11216.

Ling Xing, Alex Jinpeng Wang, Rui Yan, Xiangbo Shu, and Jinhui Tang. 2026. Vision-centric token compression in large language model. *Advances in Neural Information Processing Systems*, 38:33080–33110.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *Preprint*, arXiv:2410.10594.

Ziyang Zeng, Dun Zhang, Jiacheng Li, Panxiang Zou, Yudong Zhou, and Yuqing Yang. 2025. An empirical study of position bias in modern information retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5069–5081.

Ziyang Zeng, Dun Zhang, Yu Yan, Xu Sun, Cuiqiaoshu Pan, Yudong Zhou, and Yuqing Yang. 2026. Posir: Position-aware heterogeneous information retrieval benchmark. *Preprint*, arXiv:2601.08363.

Hongbo Zhao, Meng Wang, Fei Zhu, Wenzhuo Liu, Bolin Ni, Fanhu Zeng, Gaofeng Meng, and Zhaoxiang Zhang. 2025. Vtcbench: Can vision-language models understand long context with vision-text compression? *Preprint*, arXiv:2512.15649.

Jianping Zhong, Guochang Li, Chen Zhi, Junxiao Han, Zhen Qin, Xinkui Zhao, Nan Wang, Shuiguang Deng, and Jianwei Yin. 2026. Can vision-language models handle long-context code? an empirical study on visual compression. *Preprint*, arXiv:2602.00746.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. Longembed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 802–816.

A Dataset and Rendering Details

A.1 Domain Coverage

The 31 domains used in our full-scale protocol are inherited from PosIR (Zeng et al., 2026) and span a broad range of subject areas. Table 5 summarizes per-domain statistics under baseline rendering.

A.2 Rendering Pipeline

Each PosIR document is rendered into a plain-text image using the Python Imaging Library (PIL). The baseline configuration follows the description in §3: image width is fixed at 1000 px while height grows with content, Arial 24 px is used as the default typeface, and lines are set with $1.5\times$ spacing. Plain text is rendered without headers, figures, tables, columns, or page numbers, so that any variation across vertical positions reflects the model's positional processing rather than layout artifacts. For the rendering variant experiments in §4.3.2, only one parameter is changed at a time while the other three are held at their baseline values; this single-factor design ensures that any observed performance change can be attributed to the varied parameter alone.

We partition rendered images into four height buckets, used consistently throughout §4.3.1 and the corresponding appendix tables: *short* ($\leq$

Domain	Docs	Queries	chars	height (px)
Accommodation Catering Hotel	439,993	14,612	4188	2355
Aerospace	581,933	13,546	4283	2285
Agriculture Forestry Animal Husbandry Fishery	581,329	16,175	4334	2315
Artificial Intelligence Machine Learning	551,318	11,154	4212	2733
Automobile	565,317	13,047	4209	2224
Biomedicine	616,910	13,948	4295	2286
Computer Communication	620,254	14,144	4447	2527
Computer Programming Code	519,815	6,522	4289	3412
Current Affairs Government Administration	606,998	12,832	4482	2329
Electric Power Energy	599,389	13,119	4444	2341
Film Entertainment	596,575	16,911	4203	2211
Finance Economics	607,364	12,948	4476	2372
Fineweb	698,633	16,748	4196	2265
Fire Safety Food Safety	313,846	12,852	4468	2403
Game	442,997	13,014	4449	2452
Law Judiciary	603,544	12,947	4452	2364
Literature Emotion	598,538	15,717	4226	2266
Mathematics Statistics	618,864	10,324	3711	2684
Medicine Health Psychology Traditional Chinese Medicine	633,903	13,925	4360	2340
Mining	485,528	13,345	4096	2200
News Media	539,996	12,879	4324	2259
Other Information Services Information Security	321,187	8,590	4785	2664
Other Manufacturing	596,244	16,025	4310	2290
Petrochemical	577,348	12,803	4286	2246
Real Estate Construction	567,575	16,259	4393	2339
Sports	566,912	13,787	4125	2192
Subject Education Education	618,409	16,796	4481	2388
Technology Scientific Research	614,697	14,198	4348	2348
Tourism Geography	549,553	14,974	4150	2277
Transportation	547,312	14,129	4344	2328
Water Resources Ocean	547,392	13,438	4303	2304
Total / mean	**17,329,673**	**421,708**	**4304**	**2379**

Table 5: Per-domain statistics under baseline rendering. Docs and Queries correspond to corpus and query counts, respectively. Character length and rendered image height report rounded mean values.

1024 px), *medium* (1024–2048 px), *tall* (2048–3072 px), and *very tall* ($>$ 3072 px). The empirical relationship between source character count and rendered image height under this pipeline is detailed in §A.3.

A.3 Length–Height Correspondence

A central methodological point in §4.3.1 is that the text-retriever length buckets and the visual-retriever height buckets are defined on the same set of instances: we sort documents by character count and choose thresholds such that each text bucket contains exactly the same instances as the corresponding visual height bucket. Table 6 reports the empirical correspondence between rendered image height and source character count under baseline rendering.

The mapping is approximately linear by construction. Under fixed width (1000 px), fixed font size (24 px), and fixed line spacing (1.5×), the number of characters per line and the number of lines per vertical pixel are both content-independent up to line-wrapping fluctuation at sub-bucket scale. This is what justifies the bucketing argument in §3—spatial position on the rendered image is monotonic with textual position in the source—and what makes the visual height buckets and the text character-count buckets in Table 2 measure the *same* underlying input-length regime, so that visual–text PSI differences reflect modality processing rather than partition definitions.

A.4 Position Annotation

Position annotation follows the procedure described in §3. For each rendered image, we record the bounding box of the relevant region and assign it to a position bucket by the vertical center of that box.

Height bucket	Pixel range (px)	Char count (mean ± std)	Char range (min–max)	% of corpus
short	≤ 1024	1097 ± 340	514–1923	22.50
medium	1024–2048	2605 ± 704	1034–4202	25.26
tall	2048–3072	4697 ± 753	2540–6169	20.86
very tall	> 3072	7583 ± 1304	4975–10386	31.38

Table 6: Empirical correspondence between rendered image height and source character count under baseline rendering (1000 px width, Arial 24 px, $1.5\times$ line spacing). Height buckets are defined on rendered pixel height and are mutually exclusive, so the % of corpus column is a partition and sums to 100%. The character-count ranges, by contrast, are the empirical min–max of the instances falling into each height bucket; because content-dependent line wrapping makes the height–length mapping monotonic only up to sub-bucket fluctuation, adjacent buckets have overlapping character ranges. This overlap is expected and does not affect the partition: each visual height bucket still corresponds to exactly the same set of instances as the matched text-retriever length bucket in Table 2.

A.5 Exact Model and Inference Configurations

To facilitate reproducibility, Table 7 lists the exact checkpoints and repository identifiers used for all visual and text retrievers evaluated in our experiments.

Experiments were conducted on a server equipped with four NVIDIA RTX A6000 GPUs, using Python 3.9.25, PyTorch 2.4.0 with CUDA 12.1, and Transformers 4.45.0. Model inference used FP16 or BF16 precision depending on the model.

B Complete Full-Scale Protocol Results

B.1 NDCG@5 Results

Table 8 provides the NDCG@5 counterpart of main-paper Table 1, serving as a metric-robustness check for the F2 (shape) and F3 (magnitude) findings. Across both metrics the model ranking is preserved (within rounding), the bias direction reported in Δ_{Middle} is reproduced (positive for seven of eight visual models), and both the micro and macro PSI columns reproduce the inverted-U shape, indicating that the magnitude–shape decomposition is not an artifact of any single retrieval metric.

B.2 Per-Domain NDCG@5 and PSI

Figure 5 shows the distribution of per-domain NDCG@5 across 31 domains for each model. ColQwen achieves the highest median and the most compact distribution, while DSE-Wikiss has both the lowest median and the widest spread, with a prominent lower tail extending below 40. DSE-Docmatix and ColPali occupy intermediate positions, though ColPali's distribution is slightly more dispersed. The cross-domain gap between the best and worst domain ranges from 19.94 points (ColQwen) to 37.84 points (DSE-Docmatix), but the ordering among models is largely preserved within individual domains, suggesting that domain difficulty is primarily an intrinsic property of the content.

Figure 5 shows the corresponding distribution of per-domain PSI. DSE-Wikiss exhibits the highest median PSI and the greatest dispersion, with several domains exceeding 0.20 — consistent with its overall weaker retrieval performance amplifying positional sensitivity. DSE-Docmatix, by contrast, has the lowest median PSI and the tightest concentration, with most domains falling below 0.13. ColPali and ColQwen have similar median PSI values, but their distributions differ in shape: ColPali's mass is spread more evenly, whereas ColQwen shows a bimodal tendency with a cluster near 0.04–0.06 and another near 0.12–0.16. Across all four models, PSI varies by an order of magnitude across domains, confirming that position bias interacts with domain-specific characteristics.

B.3 Statistical Significance of Position Dependence and Middle Advantage

We assess the statistical significance of the middle advantage and overall position dependence.

Middle advantage. For each model, we resampled its available domains 10,000 times and recomputed Δ_{Middle} as macro Middle minus the higher of macro Top and macro Bottom. The three position scores from each domain were resampled together. We report the 2.5th and 97.5th percentiles as a 95% bootstrap confidence interval. Middle is numerically highest for seven of eight visual retrievers. Three models show a statistically reliable advantage, while the other four show a weaker middle-leaning trend with greater uncertainty at the

Model	Checkpoint / repository ID
DSE-Docmatix	Tevatron/dse-phi3-docmatix-v2
DSE-Wikiss	Tevatron/dse-phi3-v1.0
ColPali	vidore/colpali-v1.1
ColQwen	vidore/colqwen2-v1.0
ColNomic	nomic-ai/colnomic-embed-multimodal-3b
ColJinaV4	jinaai/jina-embeddings-v4
NomicEmbed	nomic-ai/nomic-embed-multimodal-3b
JinaV4	jinaai/jina-embeddings-v4
NomicEmbed-text	nomic-ai/nomic-embed-text-v1.5
JinaV4-text	jinaai/jina-embeddings-v4
BGE-M3	BAAI/bge-m3
BGE-Reranker-v2-m3	BAAI/bge-reranker-v2-m3

Table 7: Exact model checkpoints and repository identifiers used in our experiments.

Model	Overall	Top	Middle	Bottom	PSI (Micro)	PSI (Macro)
Dense visual retrievers						
DSE-Docmatix	67.28±6.72	64.43±7.98	68.80±6.27	67.95±6.86	0.064	0.084
DSE-Wikiss	53.93±5.80	49.68±6.37	55.98±6.35	55.21±5.69	0.113	0.128
NomicEmbed	67.52±4.68	67.81±5.12	68.25±5.23	66.37±4.90	0.028	0.057
JinaV4	65.64±6.43	65.45±7.45	66.91±6.20	64.25±6.87	0.040	0.075
Late-interaction visual retrievers						
ColPali	63.38±5.23	62.67±5.60	66.09±5.96	60.62±5.49	0.083	0.106
ColQwen	76.00±4.23	71.56±4.84	77.60±4.39	77.99±4.68	0.082	0.093
ColNomic	87.35±3.50	86.76±3.85	87.76±3.77	87.34±3.66	0.011	0.033
ColJinaV4	84.24±4.67	83.77±4.98	85.30±4.64	83.42±5.03	0.022	0.037

Table 8: NDCG@5 under the full-scale protocol, macro-averaged across 31 domains. This table serves as a metric-robustness check for the F1 shape and F2 magnitude findings: the inverted-U pattern and the magnitude–shape decomposition reported on Recall@1 in Table 1 are also observed under NDCG@5.

individual-model level. ColQwen remains slightly Bottom-highest. Table 9 reports the model-wise confidence intervals.

Overall position dependence. We also ran a stratified randomization test under the null of position exchangeability. Within each domain, successful retrievals were randomly reassigned across positions while preserving the bucket sizes and each model's total number of successful retrievals. We recomputed macro PSI over 10,000 randomizations and applied Holm correction across the eight model-wise tests. All eight visual retrievers had significantly greater PSI than their random-position null distributions, with adjusted p-values ≤ 0.0014 (Table 9). Thus, position dependence is statistically supported across all eight models, whereas a statistically reliable middle advantage appears for a subset.

B.4 Controlled Position-Shift Experiment

We conducted a position-shift control experiment. We placed the same relevant span at Top, Middle, and Bottom while preserving the query, the remaining document text, and candidate-pool membership. Each condition used the same 1,710 queries, balanced by original evidence position within four domains, with at most one query per document. The four domains were Accommodation Catering Hotel (420 queries), Electric Power Energy (426), Literature Emotion (420), and Real Estate Construction (444). Selected document renderings were replaced within the original corpus, while unselected documents remained unchanged. ColPali and ColQwen represented the strongest middle advantage and the only non-Middle peak in Table 1, respectively. For each contrast, we ran 10,000 paired bootstrap replicates, resampling queries within domains with shared indices across the compared variants and macro-averaging domain-level differences. We report percentile 95% confidence intervals using ran-

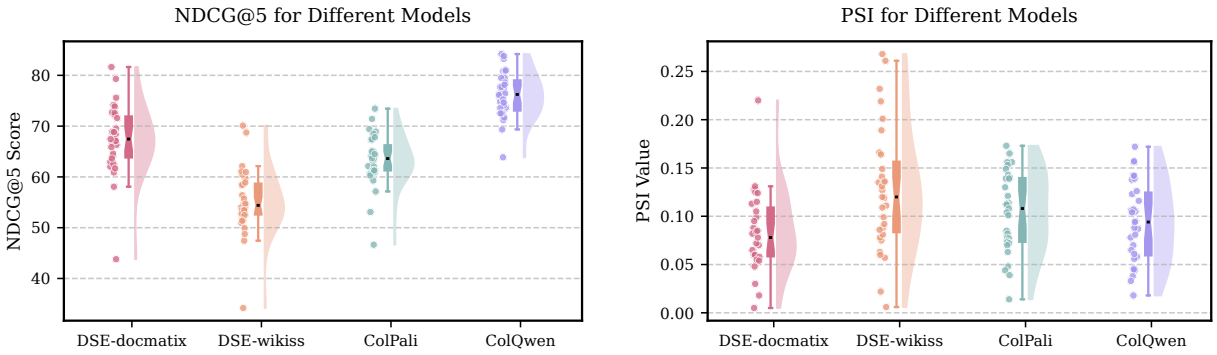


Figure 5: Raincloud plots of per-domain NDCG@5 (left) and PSI (right) across 31 domains under the full-scale protocol. Each plot combines a half-violin density estimate, a box plot, and individual domain data points.

Model	Δ_{Middle} (pp), 95% CI	Observed macro PSI	Null 95% range	Holm-adjusted p
ColNomic	$+0.22[-1.03, 1.43]$	0.061	$[0.035, 0.051]$	0.0008
ColJinaV4	$+1.85[0.59, 2.87]$	0.064	$[0.039, 0.056]$	0.0014
ColQwen	$-0.44[-1.40, 0.52]$	0.127	$[0.048, 0.069]$	0.0008
DSE-Docmatix	$+0.67[-0.66, 1.92]$	0.115	$[0.058, 0.084]$	0.0008
NomicEmbed	$+1.27[-0.45, 2.43]$	0.091	$[0.057, 0.082]$	0.0014
JinaV4	$+1.73[0.07, 2.84]$	0.100	$[0.061, 0.087]$	0.0008
ColPali	$+3.97[2.46, 4.98]$	0.121	$[0.060, 0.088]$	0.0008
DSE-Wikiss	$+0.06[-1.10, 1.26]$	0.168	$[0.075, 0.108]$	0.0008

Table 9: Statistical analyses of the middle advantage and position dependence based on Recall@1. The confidence intervals for Δ_{Middle} are calculated per model without adjustment. The null ranges for PSI cover the central 95% of the randomization distributions, along with the corresponding p-values.

dom seed 42. Because this diagnostic covers two representative models and moving a span may alter its local discourse context, we view the experiment as supporting behavioral evidence rather than a conclusive explanation of the bias.

C Image Height Effects — Complete Results

C.1 DSE Models under Height Bucketing

Table 16 extends main-paper Table 2 with the two DSE models. Both follow the same overall pattern as ColQwen — degraded accuracy with increasing height, stable Top-weakness — but DSE-Wikiss shows a steeper degradation slope, dropping to 32.76% Recall@1 in the very-tall bucket.

DSE-Docmatix presents a noteworthy artifact: its micro PSI decreases monotonically with height (0.110 → 0.010), converging to near-identical scores across all three positions in the very_tall bucket (56.25, 56.55, 56.81). This convergence reflects uniform degradation rather than genuine robustness, as the macro PSI in the same bucket (0.131) remains an order of magnitude higher — per-domain biases persist but cancel under pooling because they point in inconsistent directions.

C.2 ColPali's Bias Reversal — Per-Domain Evidence

Table 11 counts, for each (model, height) cell, how many of the 31 domains have each position as the model's best. ColPali's reversal is clearly visible: Bot is the most common best position in the short bucket (14/31), but collapses to 1/31 in the very-tall bucket, where Top takes over (24/31). No other model exhibits a comparable directional shift. ColQwen maintains Bot as the most frequent best position in all four height regimes (16, 17, 13, 20 out of 31).

C.3 PSI vs. Image Height

Several per-domain cases exhibit extreme height-induced bias amplification: ColPali on `electric_power_energy` rises from PSI 0.063 (short) to 0.634 (very_tall); ColPali on `current_affairs_government_administration` from 0.084 to 0.533; and DSE-Wikiss on `current_affairs_government_administration`

Model	Metric	Top	Middle	Bottom	Middle−Top, 95% CI	Middle−Bottom, 95% CI
ColPali	R@1	60.44	60.30	51.34	$-0.14[-1.94, 1.64]$	$+8.97[7.02, 10.91]$
ColPali	NDCG@5	69.86	69.69	61.80	$-0.17[-1.50, 1.12]$	$+7.89[6.40, 9.38]$
ColQwen	R@1	64.27	63.72	63.00	$-0.54[-2.12, 0.97]$	$+0.72[-0.76, 2.20]$
ColQwen	NDCG@5	75.31	75.11	74.31	$-0.20[-1.20, 0.77]$	$+0.80[-0.17, 1.74]$

Table 10: Position-shift results on the same 1,710 queries. Scores are macro-averaged across four domains and reported on a 0–100 scale. Differences and confidence intervals are in percentage points.

Model	Height	Top	Middle	Bottom
DSE-docmatix	short	2	17	12
	medium	5	6	19
	tall	11	12	7
	very tall	8	10	12
DSE-wikiss	short	2	14	13
	medium	2	13	14
	tall	10	6	13
	very tall	8	9	12
ColPali	short	4	12	14
	medium	8	7	15
	tall	19	11	0
	very tall	24	5	1
ColQwen	short	2	12	16
	medium	4	9	17
	tall	4	13	13
	very tall	4	6	20

Table 11: Number of domains (out of 31) in which each position is the best, by height bucket.

from 0.128 to 0.572.[1] These cases reinforce the finding that height-bias interaction is most aggressive for the fixed-resolution model and in domains where retrieval is already difficult.

D BEIR Task-Boundary Analysis

This appendix supports the task-boundary analysis in §4.4. We evaluate visual and text models from the NomicEmbed and JinaV4 families on five BEIR datasets (Thakur et al., 2021), using the baseline rendering pipeline in §3. Unlike PosIR, BEIR provides document-level relevance labels rather than localized evidence spans. Table 12 shows that both text models achieve higher mean NDCG@10 and lead on four of the five datasets; ArguAna is the sole exception. To investigate this performance reversal, we examine document length and span dilution as possible explanations for the difference between PosIR and BEIR. We also consider query length, the number of relevant documents, and lexical overlap to understand why visual models perform differently on ArguAna and the other BEIR datasets. Only lexical overlap exhibits a consistent association with the reversal (§4.4), whereas the remaining four factors do not provide a consistent explanation, as discussed in Hypotheses 1–4 below.

Hypothesis 1: document length. BEIR documents are generally short, potentially limiting the benefit of visual retrieval on longer inputs. We therefore compare visual and text retrieval within rendered-height buckets (Table 13). Both text models lead in every bucket above 512 px, including the tallest; the only visual lead occurs in the shortest bucket for NomicEmbed. Thus, short document length alone does not account for the reversal.

Hypothesis 2: span dilution. In PosIR, the relevant content for a query is a short span embedded in a longer document, whereas BEIR defines relevance at the document level. Representing the full document may therefore dilute the relevant span in PosIR, potentially disadvantaging text retrieval. We test this explanation on the PosIR sampled set by replacing each corpus document with its annotated relevant span. This reduces visual Recall@1 by 4.90–6.59 points while leaving text performance nearly unchanged, and both visual models still outperform their text counterparts (Table 14). Span dilution alone therefore does not account for the reversal.

[1] These per-domain PSI values are computed from the full set of per-domain results; see the supplementary material for the complete per-domain height-bias table.

Model	ArguAna	FiQA	NFCorpus	SciDocs	SciFact	Mean
NomicEmbed (visual)	40.70	35.66	27.20	15.35	67.98	37.38
NomicEmbed-text	35.58	37.64	34.59	18.72	70.43	**39.39**
JinaV4 (visual)	39.98	37.31	36.17	18.46	73.62	41.11
JinaV4-text	39.28	47.58	38.48	21.89	76.29	**44.70**

Table 12: BEIR retrieval performance (NDCG@10) on five datasets; all corpus documents are rendered through the baseline pipeline (§3). Within each same-backbone pair, the text retriever's mean (**bold**) is no lower than that of its visual counterpart. Unlike on PosIR, visualization does not improve over text retrieval here.

Model	0–512	512–1024	1024–2048	2048+
NomicEmbed (vis.)	39.10	38.77	39.53	35.27
NomicEmbed-text	36.10	40.95	42.98	35.41
JinaV4 (vis.)	40.29	41.08	43.05	33.47
JinaV4-text	41.97	46.09	46.16	35.68

Table 13: BEIR performance (NDCG@10) stratified by rendered image height in pixels; bucket boundaries are chosen for BEIR's shorter documents. The text retriever leads in all buckets above 512 px, including the tallest. The document-length hypothesis—which predicts a visual advantage on taller images—is not supported.

Model	Full document	Span only
NomicEmbed (vis.)	89.85 ± 0.43	83.26 ± 0.36
NomicEmbed-text	78.45 ± 0.35	78.72 ± 0.43
JinaV4 (vis.)	88.93 ± 0.45	84.03 ± 0.26
JinaV4-text	81.65 ± 0.28	81.65 ± 0.28

Table 14: Span-dilution ablation on the PosIR sampled set (Recall@1). *Full document*: the standard setting, where the relevant span is embedded in the full corpus document. *Span only*: each corpus document is replaced by its relevance span. The visual retriever still leads its text counterpart under the span-only setting, so span dilution does not explain the BEIR gap.

Hypothesis 3: query length. On the five BEIR datasets, both text models achieve higher mean NDCG@10, while visual models lead only on ArguAna (Table 12). ArguAna queries are longer than those in the other datasets, raising the possibility that query length explains its visual advantage. The observed ranges do not overlap: ArguAna queries contain 45–864 words, compared with 1–31 words elsewhere. We therefore examine query-length groups within ArguAna. Longer queries do not consistently yield a larger visual advantage, providing no consistent support for this explanation.

Hypothesis 4: number of relevant documents. ArguAna has one relevant document per query, which may contribute to its distinct modality ordering. We examine this possibility within FiQA, where the number of relevant documents varies. Both text models still lead for queries with a single relevant document, by 2.50 points for NomicEmbed and 10.04 points for JinaV4 in NDCG@10. The modality gap also varies nonmonotonically across relevance-count groups. These results do not support relevance count alone as an explanation. Table 15 reports the grouped results for Hypotheses 3 and 4.

Group	Queries	Nomic gap	Jina gap
(a) ArguAna: query length (words)			
45–133	350	+5.30	+1.43
134–175	354	+4.49	+0.56
176–230	345	+4.78	+0.63
231–864	352	+5.96	+0.19
(b) FiQA: relevant documents per query			
1	220	−2.50	−10.04
2–3	287	−0.91	−10.08
4–5	85	−2.48	−11.09
6–10	50	−4.36	−10.68
11+	6	−7.75	−12.73

Table 15: Visual-minus-text NDCG@10 differences in percentage points, grouped by query length within ArguAna and by the number of relevant documents per query within FiQA. Positive values favor visual retrieval; negative values favor text retrieval.

E Rendering Parameter Effects — Complete Results

All results in this section use the sampled protocol (§3): 10 randomly selected domains, 300 stratified samples per position bucket per domain, repeated

Model	Height	Overall	Top	Middle	Bottom	PSI_μ	PSI_M
Text retrievers (length-bucketed to match visual height buckets)							
BGE-M3	short	71.75	73.23	72.15	69.87	0.046	0.109
	medium	65.49	72.64	63.51	60.33	0.169	0.178
	tall	57.05	63.42	55.58	52.15	0.178	0.213
	very tall	52.62	57.56	49.13	51.18	0.146	0.174
NomicEmbed-text	short	70.22	71.65	70.88	68.14	0.049	0.117
	medium	62.65	69.25	61.69	57.00	0.177	0.187
	tall	49.07	58.57	48.00	40.63	0.306	0.315
	very tall	38.81	49.02	34.67	32.76	0.332	0.361
JinaV4-text	short	73.03	76.09	72.98	70.04	0.079	0.139
	medium	64.61	69.48	62.79	61.55	0.114	0.142
	tall	53.68	58.85	53.51	48.69	0.173	0.214
	very tall	49.35	54.87	47.56	45.62	0.169	0.205
BGE-Reranker-v2	short	87.97	85.27	88.17	90.47	0.057	0.081
	medium	83.58	91.18	88.44	71.12	0.220	0.222
	tall	66.96	89.70	66.40	44.79	0.501	0.501
	very tall	55.85	84.47	43.71	39.36	0.534	0.550
Visual retrievers (height-bucketed in pixels)							
DSE-Docmatix	short	74.02	68.40	76.87	76.78	0.110	0.145
	medium	72.87	71.00	73.35	74.28	0.044	0.081
	tall	66.39	66.62	67.50	65.04	0.036	0.111
	very tall	56.54	56.25	56.55	56.81	0.010	0.131
DSE-Wikiss	short	71.59	67.58	73.28	73.92	0.086	0.123
	medium	64.18	62.26	65.01	65.28	0.046	0.104
	tall	50.52	49.37	50.97	51.24	0.037	0.163
	very tall	32.76	31.16	32.69	34.44	0.095	0.199
ColPali	short	80.66	77.74	81.67	82.57	0.059	0.098
	medium	79.89	79.17	80.18	80.32	0.014	0.059
	tall	62.42	67.87	66.38	53.00	0.219	0.240
	very tall	34.11	39.57	33.90	28.85	0.271	0.296
ColQwen	short	87.31	84.81	88.16	88.95	0.047	0.078
	medium	86.08	84.55	86.62	87.08	0.029	0.054
	tall	79.31	76.80	80.08	81.06	0.053	0.083
	very tall	54.45	51.11	54.42	57.83	0.116	0.156
ColJinaV4	short	89.90	87.50	91.44	90.76	0.043	0.063
	medium	87.44	86.97	87.97	87.37	0.011	0.047
	tall	83.44	84.11	84.04	82.18	0.023	0.070
	very tall	76.96	78.70	77.19	75.00	0.047	0.084
ColNomic	short	90.50	89.51	90.84	91.17	0.018	0.049
	medium	88.96	88.36	89.20	89.32	0.011	0.051
	tall	86.70	86.19	86.54	87.37	0.014	0.061
	very tall	83.29	84.09	83.63	82.15	0.023	0.061
JinaV4	short	75.98	71.12	77.97	78.86	0.098	0.131
	medium	70.44	71.31	71.21	68.80	0.035	0.066
	tall	60.80	63.59	60.93	57.87	0.090	0.133
	very tall	56.18	58.78	55.81	53.95	0.082	0.142
NomicEmbed	short	77.77	75.44	78.79	79.09	0.046	0.088
	medium	72.02	71.69	71.74	72.63	0.013	0.080
	tall	64.24	66.75	65.12	60.86	0.088	0.135
	very tall	57.20	60.93	56.18	54.49	0.106	0.157

Table 16: Recall@1 and PSI by input length for all evaluated retrievers under the full-scale protocol. Text retrievers are length-bucketed by character count; visual retrievers are height-bucketed in pixels; the two bucketings are aligned to the same instances (§A.3). PSI_μ and PSI_M denote micro and macro PSI. Within every text retriever, Top is the highest-scoring position in all four buckets, i.e. no directional reversal; among visual retrievers, only ColPali, ColNomic, JinaV4, NomicEmbed, and ColJinaV4 reverse direction at large heights.

over 5 random seeds. ★ marks the baseline configuration.

For the three fixed-resolution models, both overall accuracy and PSI degrade substantially as font size increases, with PSI rising up to 21-fold for DSE-Docmatix (0.008 at 24 px → 0.169 at 48 px). ColQwen, the only dynamic-resolution model, shows the opposite trend on PSI (0.029 → 0.015) and a slight gain in overall accuracy (93.78 → 95.12).

E.1 Line Spacing

Across the 1.0×–2.0× range, PSI changes by a factor of 2–4 for the fixed-resolution models, far less than the 21-fold change driven by font size, consistent with the main-paper observation that line spacing has the weakest effect on position bias among the three rendering parameters.

E.2 Typeface

Familiar typefaces (Arial, Times New Roman) yield the strongest performance across all models, likely because they are well represented in the pre-training data of the underlying vision-language backbones. Courier New induces disproportionate degradation in the three fixed-resolution models — DSE-Wikiss drops from 81.72 to 68.40, ColPali from 83.90 to 65.60 — while ColQwen is comparatively robust (93.68 → 91.68). Courier New is the only monospace typeface in our set: all glyphs occupy the same horizontal advance width regardless of character shape, producing sparse pixel coverage for narrow characters such as 'i' and 'l' while wide characters retain normal density. The resulting reduction in average per-word pixel density appears to interact poorly with fixed-patch encoders that allocate a fixed budget of patches per pixel. Handwriting-style fonts, despite their visual irregularity, do not cause comparable degradation, suggesting that it is the pixel-density structure of Courier New rather than glyph irregularity *per se* that drives the performance drop.

E.3 Height-Stratified Font-Size Analysis

We compare font sizes within the same rendered-height buckets: Short ($\leq$ 1,024 px), Medium (1,024–2,048 px), Tall (2,048–3,072 px), and Very tall ($>$ 3,072 px). Table 20 reports R@1 and PSI averaged over available seed-level estimates. PSI is computed within each run before averaging. The results suggest a resolution–context trade-off: small glyphs may be less legible within visual patches, whereas larger glyphs spread the same text over a taller image. We present this as an interpretation, not a verified mechanism, because documents within a height bucket can differ in source length and sample composition across font sizes.

F Gradient Saliency Analysis

We differentiate the native MaxSim query–positive-document score with respect to visual-token hidden states. Saliency is the L_2 norm of the gradient for each valid spatial token, normalized by the sum over the page. Tokens are mapped to page coordinates using the verified ColPali patch grid or the merged ColQwen visual grid. The evidence-containing region denotes a page third, not the exact evidence bounding box.

The ColPali spatial analysis uses 2,000 query–positive-document pairs. The evidence-position comparison uses gradients at the final decoder-block input, whereas the height analysis uses the first-block input. Figure 4(a) includes bottom-evidence examples per height bucket.

The font analysis uses the same pairs shared across the 12, 24, and 48 px renderings. Both models use gradients at the final decoder-block input, and all positive pairs are included regardless of retrieval rank. Figure 4(b) reports paired changes in evidence-region saliency. Grid dimensions and visual-token counts in the ColQwen analysis are averaged separately over the same pairs.

Model	Font size	Overall	Top	Middle	Bottom	PSI
DSE-Docmatix	12	91.24 ± 0.40	90.94 ± 0.50	92.34 ± 0.50	90.44 ± 0.55	0.021
	16	91.24 ± 0.44	90.70 ± 0.47	92.26 ± 0.40	90.74 ± 0.81	0.017
	24★	89.38 ± 0.57	89.14 ± 0.68	89.88 ± 0.41	89.12 ± 1.02	0.008
	32	76.30 ± 0.34	72.52 ± 0.34	78.88 ± 0.43	77.40 ± 0.96	0.081
	48	48.95 ± 0.83	43.35 ± 1.05	52.15 ± 0.73	51.10 ± 1.50	0.169
DSE-Wikiss	12	82.52 ± 0.83	82.32 ± 0.99	84.12 ± 0.76	81.16 ± 0.96	0.035
	16	83.04 ± 1.02	82.10 ± 1.04	84.56 ± 1.06	82.48 ± 1.15	0.013
	24★	81.72 ± 0.79	80.12 ± 0.75	83.22 ± 0.95	81.78 ± 1.13	0.038
	32	66.64 ± 0.17	61.80 ± 0.96	69.82 ± 0.77	67.84 ± 0.39	0.115
	48	42.50 ± 0.86	36.62 ± 1.00	45.55 ± 0.83	45.12 ± 1.66	0.196
ColPali	12	88.06 ± 0.58	88.14 ± 0.57	91.46 ± 0.90	84.52 ± 0.61	0.076
	16	87.64 ± 0.65	88.70 ± 0.87	89.34 ± 3.16	85.28 ± 3.15	0.045
	24★	82.96 ± 0.70	83.96 ± 0.74	85.02 ± 0.45	79.92 ± 1.18	0.060
	32	70.02 ± 0.24	67.78 ± 0.63	72.88 ± 0.54	69.40 ± 0.41	0.070
	48	45.40 ± 0.94	40.92 ± 1.20	48.36 ± 0.85	46.64 ± 1.19	0.154
ColQwen	12	92.76 ± 0.24	91.04 ± 0.46	94.22 ± 0.30	92.94 ± 0.34	0.034
	16	94.08 ± 0.26	92.58 ± 0.30	95.02 ± 0.23	94.40 ± 0.42	0.026
	24★	93.78 ± 0.45	92.06 ± 0.62	94.44 ± 0.29	94.78 ± 0.59	0.029
	32	94.66 ± 0.21	93.26 ± 0.36	95.46 ± 0.25	95.54 ± 0.23	0.024
	48	95.12 ± 0.24	94.33 ± 0.13	95.72 ± 0.30	95.35 ± 0.34	0.015

Table 17: Recall@1 by font size (sampled protocol). The baseline 24 px configuration is marked with ★.

Model	Line spacing	Overall	Top	Middle	Bottom	PSI
DSE-Docmatix	1.0	90.38 ± 0.45	90.22 ± 0.48	91.08 ± 0.75	89.82 ± 0.47	0.014
	1.25	90.42 ± 0.61	90.04 ± 0.61	91.06 ± 0.67	90.16 ± 0.65	0.011
	1.5★	89.38 ± 0.57	89.14 ± 0.68	89.88 ± 0.41	89.12 ± 1.02	0.008
	1.75	88.35 ± 0.57	87.45 ± 0.83	89.33 ± 0.38	88.23 ± 0.60	0.021
	2.0	86.40 ± 0.54	85.05 ± 0.78	87.88 ± 0.64	86.28 ± 0.45	0.032
DSE-Wikiss	1.0	81.82 ± 0.86	80.26 ± 0.98	82.88 ± 0.94	82.26 ± 1.05	0.032
	1.25	80.20 ± 0.65	78.48 ± 0.74	81.54 ± 0.54	80.54 ± 1.26	0.038
	1.5★	81.72 ± 0.79	80.12 ± 0.75	83.22 ± 0.95	81.78 ± 1.13	0.038
	1.75	78.05 ± 0.73	75.65 ± 0.98	79.67 ± 0.50	78.75 ± 0.90	0.050
	2.0	75.03 ± 0.54	72.38 ± 0.90	77.28 ± 0.35	75.48 ± 0.63	0.063
ColPali	1.0	83.90 ± 0.65	84.64 ± 0.74	86.06 ± 0.72	81.12 ± 0.63	0.057
	1.25	81.16 ± 0.40	81.78 ± 0.55	83.64 ± 0.57	78.16 ± 0.49	0.066
	1.5★	82.96 ± 0.70	83.96 ± 0.74	85.02 ± 0.45	79.92 ± 1.18	0.060
	1.75	81.16 ± 0.40	81.78 ± 0.55	83.64 ± 0.57	78.16 ± 0.49	0.066
	2.0	79.04 ± 0.44	79.20 ± 0.78	81.68 ± 0.76	76.30 ± 0.37	0.066
ColQwen	1.0	93.68 ± 0.40	92.04 ± 0.44	94.48 ± 0.24	94.48 ± 0.54	0.026
	1.25	93.52 ± 0.38	91.98 ± 0.62	94.32 ± 0.57	94.20 ± 0.68	0.025
	1.5★	93.78 ± 0.45	92.06 ± 0.62	94.44 ± 0.29	94.78 ± 0.59	0.029
	1.75	94.52 ± 0.33	93.20 ± 0.44	95.10 ± 0.37	95.16 ± 0.49	0.021
	2.0	94.78 ± 0.26	93.66 ± 0.24	95.28 ± 0.41	95.36 ± 0.44	0.018

Table 18: Recall@1 by line spacing (sampled protocol). Baseline is 1.5×, marked with ★.

Model	Typeface	Overall	Top	Middle	Bottom	PSI
DSE-Docmatix	Arial★	89.38 ± 0.57	89.14 ± 0.68	89.88 ± 0.41	89.12 ± 1.02	0.008
	Times New Roman	89.86 ± 0.50	88.68 ± 0.86	91.28 ± 0.69	89.62 ± 0.33	0.028
	Courier New	79.44 ± 0.40	76.92 ± 0.59	81.82 ± 0.49	79.50 ± 0.81	0.034
	Handwriting Comic	88.20 ± 0.59	86.85 ± 0.51	89.90 ± 0.98	87.88 ± 0.38	0.034
	DejaVu Serif	85.20 ± 0.50	83.44 ± 0.45	87.36 ± 0.77	84.72 ± 0.46	0.045
DSE-Wikiss	Arial★	81.72 ± 0.79	80.12 ± 0.75	83.22 ± 0.95	81.78 ± 1.13	0.032
	Times New Roman	80.84 ± 0.99	78.68 ± 0.83	82.40 ± 1.25	81.36 ± 1.03	0.045
	Courier New	68.40 ± 0.39	64.24 ± 0.87	71.46 ± 0.77	69.56 ± 1.27	0.101
	Handwriting Comic	78.22 ± 0.99	75.30 ± 0.80	80.70 ± 1.55	78.55 ± 0.88	0.067
	DejaVu Serif	75.10 ± 0.34	71.94 ± 0.55	77.72 ± 1.00	75.78 ± 0.49	0.074
ColPali	Arial★	83.90 ± 0.65	84.64 ± 0.74	86.06 ± 0.72	81.12 ± 0.63	0.057
	Times New Roman	82.36 ± 0.43	83.10 ± 0.79	84.94 ± 0.50	79.10 ± 0.75	0.069
	Courier New	65.60 ± 0.57	63.18 ± 0.90	69.08 ± 0.43	64.56 ± 0.85	0.085
	Handwriting Comic	80.10 ± 0.57	80.10 ± 0.84	82.50 ± 1.03	77.80 ± 1.04	0.057
	DejaVu Serif	75.02 ± 0.11	73.30 ± 0.45	78.36 ± 0.80	73.44 ± 0.78	0.065
ColQwen	Arial★	93.68 ± 0.40	92.04 ± 0.44	94.48 ± 0.24	94.48 ± 0.54	0.026
	Times New Roman	94.30 ± 0.31	92.64 ± 0.51	95.24 ± 0.40	94.96 ± 0.53	0.027
	Courier New	91.68 ± 0.43	89.66 ± 0.57	92.50 ± 0.57	92.86 ± 0.43	0.034
	Handwriting Comic	93.04 ± 0.34	91.48 ± 0.54	93.84 ± 0.48	93.70 ± 0.54	0.025
	DejaVu Serif	94.24 ± 0.23	92.84 ± 0.39	95.14 ± 0.05	94.76 ± 0.42	0.024

Table 19: Recall@1 by typeface (sampled protocol). Baseline is Arial, marked with ★.

Model	Height	12 px	16 px	24 px	32 px	48 px
ColNomic	Short	98.16/0.006	98.84/0.004	99.21/0.006	99.29/0.009	97.21/0.049
	Medium	97.53/0.010	98.57/0.008	98.94/0.005	99.25/0.005	99.23/0.009
	Tall	97.79/0.040	97.38/0.011	98.93/0.008	98.81/0.006	99.16/0.007
	Very tall	82.53/0.292	98.43/0.031	97.90/0.014	98.41/0.006	98.53/0.006
ColQwen	Short	97.09/0.006	98.13/0.007	97.96/0.017	97.88/0.028	94.28/0.067
	Medium	82.64/0.042	92.63/0.028	97.63/0.006	98.28/0.010	97.79/0.021
	Tall	72.03/0.222	83.79/0.062	97.47/0.018	98.41/0.010	98.44/0.014
	Very tall	32.52/0.503	77.16/0.161	83.71/0.062	92.28/0.033	94.50/0.015
DSE-Wikiss	Short	84.00/0.033	87.48/0.016	92.32/0.022	93.30/0.037	96.72/0.081
	Medium	79.03/0.027	80.33/0.046	87.96/0.019	92.87/0.036	96.53/0.022
	Tall	83.64/0.147	73.64/0.075	80.66/0.048	88.60/0.025	94.63/0.031
	Very tall	38.70/0.705	66.77/0.193	65.25/0.051	48.28/0.110	31.76/0.178
NomicEmbed	Short	90.75/0.027	93.36/0.013	95.01/0.031	96.17/0.017	97.94/0.057
	Medium	82.42/0.082	86.82/0.061	92.24/0.026	94.69/0.018	96.31/0.021
	Tall	81.91/0.143	79.36/0.079	89.79/0.054	92.68/0.023	94.92/0.025
	Very tall	64.34/0.334	80.56/0.154	84.07/0.073	86.62/0.056	88.62/0.043

Table 20: Recall@1 / PSI by font size within rendered-height buckets. Values are averaged over available seed-level estimates; PSI is computed within each run before averaging. Recall@1 is reported on a 0–100 scale. Documents in the same height bucket may differ across font sizes.