Text-Conditioned Background Generation for Editable Multi-Layer Documents

Taewon Kang[1*], Joseph K J[2], Chris Tensmeyer[2], Jihyung Kil[2], Wanrong Zhu[2], Ming C. Lin[1], and Vlad I. Morariu[2]

[1] University of Maryland at College Park, United States
[2] Adobe Research, United States
taewon@umd.edu, {josephkj,tensmeye,jkil,wzhu}@adobe.com, lin@umd.edu, morariu@adobe.com

Abstract. We present a framework for document-centric background generation with multi-page editing and thematic continuity. To ensure text regions remain readable, we employ a *latent masking* formulation that softly attenuates updates in the diffusion space, inspired by smooth barrier functions in physics and numerical optimization. In addition, we introduce *Automated Readability Optimization (ARO)*, which automatically places semi-transparent, rounded backing shapes behind text regions. ARO determines the minimal opacity needed to satisfy perceptual contrast standards (WCAG 2.2) relative to the underlying background, ensuring readability while maintaining aesthetic harmony without human intervention. Multi-page consistency is maintained through a summarization-and-instruction process, where each page is distilled into a compact representation that recursively guides subsequent generations. This design reflects how humans build continuity by retaining prior context, ensuring that visual motifs evolve coherently across an entire document. Our method further treats a document as a structured composition in which text, figures, and backgrounds are preserved or regenerated as separate layers, allowing targeted background editing without compromising readability. Finally, user-provided prompts allow stylistic adjustments in color and texture, balancing automated consistency with flexible customization. Our training-free framework produces visually coherent, text-preserving, and thematically aligned documents, bridging generative modeling with natural design workflows.

Keywords: Document Editing · Diffusion Models · Readability Optimization · Multi-Page Consistency · Layer-Aware Generation

1 Introduction

Designing and editing complex documents that interleave text and images—such as academic reports, educational materials, or presentation slides—remains a longstanding challenge in generative modeling. While diffusion models have made significant progress in text-to-image synthesis and style transfer, they are typically optimized for producing standalone images rather than for documents.

* Work done while first author was an Adobe Research intern and then continued as a collaborative effort with UMD.

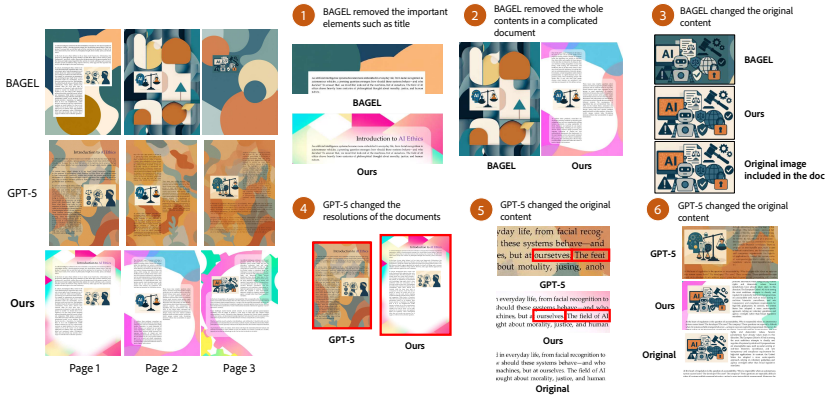

Fig. 1: Comparison with existing diffusion methods. Baseline diffusion models overwrite or alter the original document: removing titles and figures ((1), (2)), modifying semantic content ((3),(5),(6)), and even changing resolution ((4)). In contrast, our method preserves all foreground elements (text + images), while generating visually coherent, multi-page backgrounds aligned with the document content.

Applying diffusion techniques to documents requires not only producing visually coherent backgrounds but also ensuring text readability, layout fidelity, and consistency across multiple pages. Among these, background generation is particularly critical because visually salient backgrounds can interfere with readability if not handled properly, yet existing systems often under-emphasize this aspect, resulting in visual artifacts, results appearing disjointed across pages, or inconsistent with intended designs [13,33,34].

A particularly important aspect of this task is the ability to perform fine-grained background editing while maintaining global coherence. Designers may wish to adjust motifs, alter stylistic elements, or refine individual pages—without disrupting the overall flow of the document. Standard diffusion pipelines, however, often struggle to reconcile localized edits with global consistency, leading to cross-page inconsistencies [13,33,34], thereby hindering integration of generative models into real-world document workflows.

To address these gaps, we propose a document background design framework that adapts diffusion for layered, multi-page documents. Our system focuses on the specific challenge of background generation while preserving readability and layout. Our approach treats a document not as a flat image but as a structured multi-layer composition, where text, figures, and backgrounds can be independently preserved or regenerated. This layered representation enables selective modification of backgrounds while safeguarding textual content through layout-aware conditioning. At the same time, our framework emphasizes the novelty of multi-page consistency: each page is summarized into a compact representation, and this summary recursively [23] guides subsequent background generation through concise design instructions. This process ensures that visual motifs evolve coherently across an entire document, rather than drifting page by page (Fig 1).

A central component of this framework is a strategy called *latent masking*, which protects foreground regions such as text and figures during background

generation. Salient background objects often obscure document readability if left unchecked. A naive solution is to erase content using binary masks, but such hard removal introduces sharp boundaries and produces visually jarring results. Instead, our approach begins from the key idea of **purposefully weakening background generation** in foreground areas to ensure their preservation. We implement this through a soft, layout-aware attenuation mask applied in latent space, which reduces diffusion updates in regions containing text while allowing the background to evolve naturally around them. Without applying hard constraints to create discontinuities, we model masking as a smooth attenuation field in latent space, analogous to diffusion barriers in physics or weighting functions in numerical optimization. This formulation allows the background to evolve naturally around protected regions while minimizing artifacts at the boundaries, producing visually rich yet balanced results – supporting iterative refinement across pages.

Beyond masking, we introduce *Automated Readability Optimization (ARO)* to guarantee text legibility. Unlike prior systems that place uniform opaque patches behind text, ARO automatically determines [22] the minimal opacity of semi-transparent, rounded-corner backing shapes needed to satisfy perceptual contrast standards (WCAG 2.2). By analyzing the luminance distribution of the background and blending it with the overlay color, ARO computes an opacity value α that achieves a target contrast ratio across a specified coverage fraction of pixels. This ensures that inserted shapes remain visually harmonious with the background, while meeting accessibility-driven readability requirements without human intervention.

Our method also incorporates semantic grounding to align visuals with document themes. Document-level summaries guide the background generation process, aligning motifs with subject matter (e.g., fairness in AI ethics, molecular structures in life sciences). In addition, natural language prompts allow stylistic adjustments, enabling users to influence color palettes, textures, and other aesthetic choices. This combination of automated summarization with optional user constraints supports both consistent automation and flexible customization.

This work highlights how diffusion models can be adapted from pure image synthesis to structured, multi-page document background generation. By emphasizing readability, consistency, and controllability, our framework bridges generative modeling with the practical needs of document editing. Our contributions are three-folded:

1. **Latent Masking for Foreground-aware Backgrounds.** We introduce a layout-aware attenuation in latent space that reduces diffusion updates on text and figure regions while operating in a layer-separable document representation, yielding natural, non-intrusive backgrounds with preserved readability. Instead of enforcing hard binary constraints, our formulation models masking as a smooth attenuation field to support iterative refinement and stylistic consistency across pages with alignment between visual design and document content.

2. **Automated Readability Optimization (ARO).** We propose a contrast-driven, WCAG-guided method that computes the *minimal* opacity of semi-transparent rounded backings per text box for coverage-aware linear-light computation, guaranteeing legibility without manual tuning (see Fig. 6).
3. **LLM-based Multi-Page Consistency.** We employ a two-stage pipeline with a *Summarization Model* (to distill each page into a compact representation) and an *Instruction Generation Model* (to convert summaries into concise background design prompts), recursively carrying context across pages to maintain consistent motifs.

2 Related Work

2.1 Document Background and Poster Generation

Recent works have begun to explore the integration of generative models into background [12,14,18,20,28,42], layout [17,25,30,43,44,46,47] and poster creation [6,35,49]. BAGEL [13] enables intuitive text-to-design and design editing through natural language prompts, lowering the entry barrier for non-expert users. A notable advantage of BAGEL is its strong consistency in interactive editing: once an initial generation is produced, users can reliably make localized modifications to selected elements while preserving the rest of the layout. This makes BAGEL highly effective for poster-style editing. However, BAGEL struggles when applied to complex documents that interleave dense text, figures, and tables across multiple pages, where background synthesis often interferes with content fidelity and fails to maintain cross-page coherence. To address aesthetics, POSTA [6] proposes a modular framework that combines multimodal large language models and diffusion models to generate artistic posters. While this achieves high text accuracy and visually compelling backgrounds, it still requires explicit human intervention through prompt design and planning, restricting full automation. CreatiPoster [49] further improves fidelity by supporting editable, multi-layer compositions, surpassing existing commercial systems in accuracy and asset handling. Nonetheless, it also depends heavily on human-provided instructions and assets, thereby limiting scalability to large document collections. Across these approaches, the central limitation is a reliance on prompt-heavy or manual workflows, along with weak protection of critical document content such as tables, headers, or densely populated text blocks.

2.2 Diffusion for Design and Layout

Diffusion-based design frameworks such as BAGEL [13] enable iterative refinement of visual elements once an initial layout is produced, making them effective for poster-like media. However, when applied to documents that interleave dense text and embedded figures, two challenges emerge: (1) background synthesis frequently intrudes into foreground regions, harming readability [24,37,38,51,52], and (2) maintaining page-to-page stylistic coherence becomes difficult. Prior work on text-aware generation tackles complementary goals: TextDiffuser and TextDiffuser-2 [7,8] generate legible text using explicit position conditioning, and SAWNA [32] preserves empty negative space by injecting nonreactive noise. Yet these methods assume either controllable text rendering or blank regions;

they do not protect existing foreground content in complex document layouts. Classical readability studies [38] further show that when textured backgrounds interact with text, legibility decreases and can only be restored by global masking or frequency filtering. Beyond layout handling, most diffusion-based editing methods focus on quality maximization rather than content preservation. Mask-guided editing [2,10,21,29] restricts edits spatially, but text can still be overwritten if masks overlap. Attention-control models [4,5,16] sharpen generation by reinforcing semantics, whereas layered approaches [18,26,48] assume clean separable layers—an unrealistic assumption for dense document layouts. Spatial or training-free layout control [9,39,45] relies on strong external signals and does not inherently prevent degradation of embedded text. As recent surveys note [19], diffusion research overwhelmingly pursues sharper and more detailed outputs. In contrast, document-centric editing requires a fundamentally different objective: **to preserve strict text fidelity, the model must sometimes do *less*, not more**. Our latent masking softly attenuates diffusion updates in sensitive regions, analogous to smooth barrier functions in control and optimization [1,15,36,41], while our Automated Readability Optimization (ARO) module enforces WCAG contrast standards [40]. Together, attenuation in latent space and explicit contrast optimization provide principled protection of readability — often overlooked by prior works.

2.3 Interactive Document Editing

A parallel line of research emphasizes interactivity in editing. Existing frameworks such as POSTA [6], CreatiPoster [49], and InstructPix2Pix [3] adopt human-in-the-loop designs, relying on iterative prompt adjustments, localized editing, asset uploads, or layout preferences to guide the final result. While this improves user control, it also shifts the burden onto the designer, limiting efficiency for multi-page or large-scale editing tasks. More broadly, interactivity is not limited to prompt-based interfaces: conventional authoring tools such as PowerPoint also support interactive document editing. Beyond poster-oriented systems, controllable layout-to-image frameworks such as GLIGEN [27], LayoutDiffusion [50], and HiCo [31] extend interactivity by enabling grounded or hierarchical control over generation. However, these approaches still rely heavily on iterative user intervention and do not fully address the unique constraints of document-centric editing. Generative systems must therefore balance the strengths of automation with the flexibility of iterative human refinement. Moreover, systems like GPT-4o [33] and GPT-5 [34], when applied to document backgrounds, alter not only the visual layer but also the textual content itself, rendering them unsuitable for editing use cases. These observations underscore the gap between artistic generation and document-centric editing: robust systems must allow users to modify or regenerate backgrounds interactively while preserving strict text fidelity. This motivates methods that combine structured summarization, memory-driven instruction generation, and latent masking, thereby ensuring automation and interactivity without sacrificing robustness.

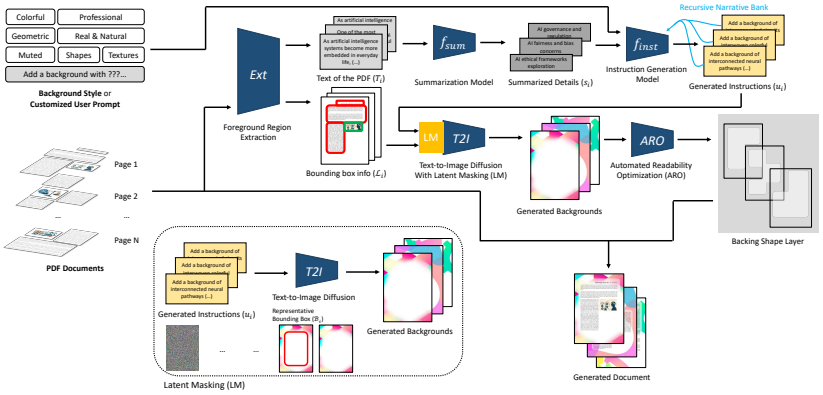

Fig. 2: Overview of our document-centric background generation framework. Given structured document pages (e.g., PDF, slides), we first perform *Foreground Region Extraction* to obtain page-level text T_i and bounding box information $\mathcal{L}_i$, while selecting representative regions $\mathcal{B}_i$ for latent masking. The *Summarization Model* compresses verbose page text T_i into a compact semantic label s_i, which is transformed into generation instructions u_i by the *Instruction Model* with multi-page continuity enforced by the Recursive Narrative Bank (RNB). Backgrounds are then synthesized by a text-to-image diffusion model, guided by (i) *Latent Masking (LM)* using $\mathcal{B}_i$ to preserve foreground readability, and (ii) *Automated Readability Optimization (ARO)* which adaptively places semi-transparent backing shapes behind all text regions $\mathcal{L}_i$ to satisfy WCAG contrast requirements. The resulting backgrounds are composited with the document foreground, yielding coherent, readable, and visually consistent multi-page documents.

3 Method

3.1 Foreground Region Extraction

Before summarization, we identify representative text regions that define the document foreground. For each page i, we detect text lines $\mathcal{L}_i = \{(\mathbf{b}_{i,\ell}, t_{i,\ell})\}$ with bounding boxes $\mathbf{b}_{i,\ell} = (x_0, y_0, x_1, y_1)$. Paragraphs are formed by grouping adjacent lines based on consistent left margins and bounded vertical gaps. Each paragraph p then receives a box $\mathbf{b}_p$ obtained as the union of its constituent line boxes.

To prevent merging across large figures, detected image zones $\mathcal{I}_i$ partition the page vertically into top, side, and bottom groups. Within each group, paragraphs are merged into column-like regions if their horizontal overlap

$$\text{overlap}_x(\mathbf{b}_p, \mathbf{b}_r) \;=\; \frac{\max\{0,\, \min(x_1^p, x_1^r) - \max(x_0^p, x_0^r)\}}{\max\{1,\, \min(x_1^p - x_0^p,\; x_1^r - x_0^r)\}} \tag{1}$$

exceeds a threshold η_x, and their vertical gap is within tolerance. Finally, an NMS-like suppression removes redundant overlaps: a candidate region p is discarded if it is largely contained in a larger region q, i.e.,

$$\frac{\text{area}(p \cap q)}{\min\{\text{area}(p), \text{area}(q)\}} \geq \tau_{\text{cont}} \quad \text{or} \quad \text{IoU}(p, q) \geq \tau_{\text{iou}}. \tag{2}$$

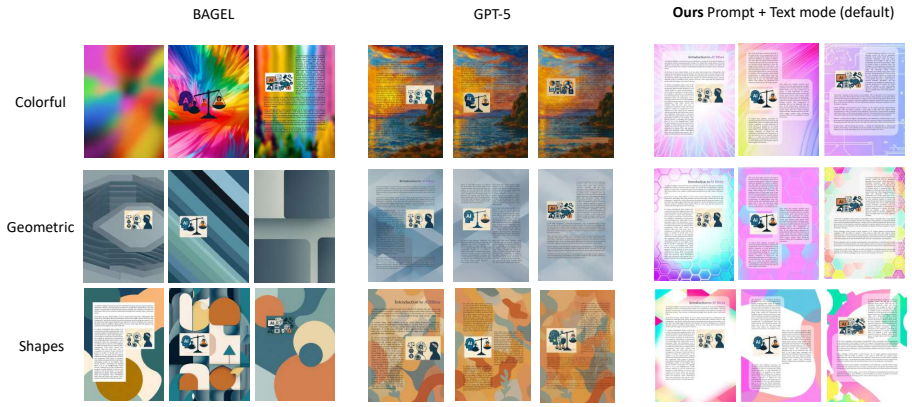

Fig. 3: Representative qualitative comparison on academic-style **PDFs** (A4). Rows correspond to style conditions (Colorful, Geometric, Muted, Professional, Real & Natural, Shapes, Textures). See more results in the supplementary materials.

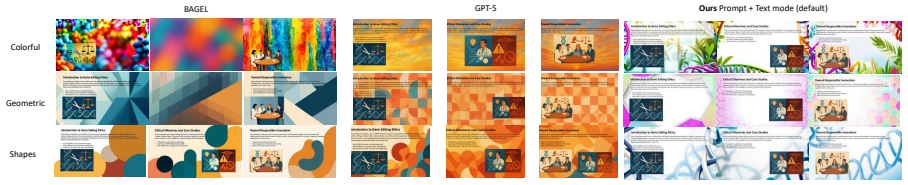

Fig. 4: Representative qualitative comparison on academic-style **slides** (16:9). Rows correspond to style conditions (Colorful, Geometric, Muted, Professional, Real & Natural, Shapes, Textures). See more results in the supplementary materials.

The surviving set $\mathcal{R}_i = \{(\mathbf{b}_r, t_r)\}$ yields representative bounding boxes $\mathcal{B}_i = \{\mathbf{b}_r\}$, while the full-page text is

$$T_i = \mathrm{concat}\big(\{t_{i,\ell} \mid (\mathbf{b}_{i,\ell}, t_{i,\ell}) \in \mathcal{L}_i\}\big). \tag{3}$$

3.2 Summarization Model

To ground visual design in document content, we first extract raw text T_i from each page i of a *structured document instance* (e.g., PDF page, slide canvas, or equivalent container), abstracting away from any specific file format. Document pages often contain multiple paragraphs with dozens of sentences, making T_i verbose and noisy for visual grounding. Directly feeding such heavy text into subsequent modules introduces two problems: (i) generation of overly complex or conflicting background prompts, and (ii) loss of robustness due to long, detailed paragraphs across many densely populated pages. We map each T_i into a compact semantic label s_i:

$$s_i = f_{\mathrm{sum}}(T_i), \tag{4}$$

where f_{sum} outputs a short phrase (five words or fewer) capturing the dominant visual theme of the page. This plays a dual role: compressing dense text into a concise representation for visual grounding, and providing semantic cues for background generation. Foreground preservation itself is handled separately using the representative bounding boxes $\mathcal{B}_i$ when constructing the latent mask M (Sec. 3.4).

Users often prefer backgrounds that are simple yet reflective of the document's theme, without specifying a prompt for every page [43]. By extracting s_i automatically, the system synthesizes backgrounds even when no user prompt p is provided. This enables prompt-free, content-driven generation that is both minimal and document-specific [11].

3.3 Instruction Generation Model

Given the semantic summary s_i and/or a user prompt p, together with the context of prior pages $H_{i-1} = \{u_1, \ldots, u_{i-1}\}$, we generate a page-level instruction u_i:

$$u_i = f_{\text{inst}}(s_i, p, H_{i-1}), \tag{5}$$

where f_{inst} produces a concise instruction describing a visual motif. Conditioning on H_{i-1} allows—but does not strictly enforce—multi-page coherence by making style cues available to subsequent pages. Unlike video or dialogue, document backgrounds must remain visually coherent across pages while reflecting local content. Stateless prompting fails here, since independently generated instructions diverge stylistically. To mitigate this, we adapt the idea of a *Recursive Narrative Bank* [23] for documents, where a structured memory of prior page-level instructions conditions subsequent generations. We define the recursive memory at step i as:

$$H_i = \{u_i^{(1)}, u_i^{(2)}, \ldots, u_i^{(N)}\}, \tag{6}$$

where $u_i^{(k)}$ denotes prior instructions within a memory window of size N. The next instruction is generated as

$$u_i = f_{\text{inst}}\big(s_i, p, H_{i-1}\big). \tag{7}$$

This recursive structure allows accumulated stylistic cues (color tones, textures) to persist across the document. Visual coherence here denotes consistent reuse of global stylistic elements (palette, tone, motifs), while still allowing local variation tied to s_i.

3.4 Foreground-Aware Latent Masking for Document Generation

We first define the vanilla diffusion update. Let x_t denote the latent state at time t and v_t^{raw} the model-predicted velocity:

$$x_{t-\Delta t} \;=\; x_t \;-\; v_t^{\text{raw}}\,\Delta t. \tag{8}$$

Mask Construction. The latent is arranged on a 2D lattice with height h and width w. We define

$$M_{ij} = \begin{cases} \lambda, & (i,j) \in \mathcal{C}(\rho; h, w), \\ 1, & \text{otherwise}, \end{cases} \tag{9}$$

where $\mathcal{C}(\rho; h, w)$ is a centered window covering a fraction ρ of the lattice; $\lambda \in (0, 1)$ is the attenuation factor. Mapping M onto token positions yields a mask $\mathbf{m}$.

Time-Gated Modulation. Masking is applied at later timesteps. The effective velocity becomes

$$v'_t = \mathbf{m} \odot v_t^{\text{raw}} + (1 - \mathbf{m}) \odot \text{stopgrad}(v_t^{\text{raw}}), \tag{10}$$

and the update is

$$x_{t-\Delta t} = x_t - v'_t \, \Delta t. \tag{11}$$

This softly attenuates generation in text regions while keeping background areas rich and variable.

Discussion. Foreground boxes (from layout analysis) provide precise regions for ARO, while representative layout boxes define the latent mask window. Related works on diffusion-based inpainting also apply region masks, but our time-gated, attenuation-based variant using Eqn. 10 and Eqn. 11 is tailored for document readability. Unlike amplification-based strategies, we deliberately suppress foreground updates, stabilizing text regions while letting the background evolve.

3.5 Automated Readability Optimization (ARO)

Instead of adding uniform opaque backing shapes to ensure readability, ARO computes the minimal opacity α^* of semi-transparent backings that meets WCAG 2.2 [40] contrast.

Contrast Calculation. Convert sRGB to linear RGB:

$$C_{\text{lin}} = \begin{cases} C_{\text{srgb}}/12.92, & C_{\text{srgb}} \leq 0.04045, \\ \big((C_{\text{srgb}} + 0.055)/1.055\big)^{2.4}, & \text{otherwise}. \end{cases} \tag{12}$$

Relative luminance:

$$L(R, G, B) = 0.2126 R_{\text{lin}} + 0.7152 G_{\text{lin}} + 0.0722 B_{\text{lin}}. \tag{13}$$

WCAG contrast:

$$CR(L_1, L_2) = \frac{\max(L_1, L_2) + 0.05}{\min(L_1, L_2) + 0.05}. \tag{14}$$

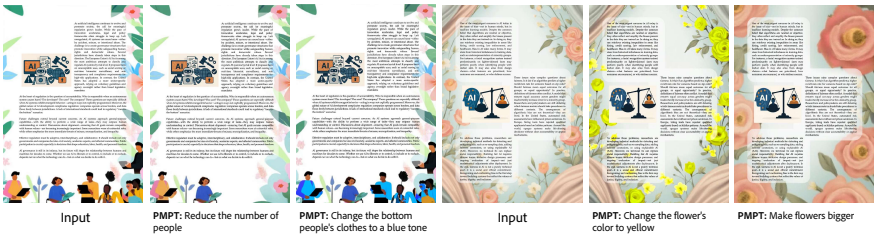

Fig. 5: Feedback-based document editing. Our system enables post-generation refinement through prompts. Users can modify only the background layer—without altering text or figures— such as reducing the number of people, adjusting colors, style and scale.

Opacity Search. For overlay luminance L_o and background pixel L_{bg}, blended luminance:

$$L_{\text{blend}}(\alpha) = \alpha L_o + (1-\alpha)L_{\text{bg}}. \tag{15}$$

Minimal opacity α^* is

$$\alpha^* = \min\left\{\alpha \,\middle|\, \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\Big[CR\big(L^{(i)}_{\text{blend}}(\alpha), L_t\big) \geq \tau\Big] \geq \rho\right\}. \tag{16}$$

Final opacity:

$$\alpha = \min\big(1,\, \max(\alpha^* + \epsilon,\, \alpha_{\min})\big). \tag{17}$$

Overlay Construction. Each bounding box $(\mathbf{b}_{i,\ell}) \in \mathcal{L}_i$ is expanded and drawn as a rounded rectangle with adaptive overlay color and computed opacity. This RGBA overlay is composited above the background.

Discussion. ARO ensures readability with minimal intervention. Unlike prior opaque patches, it adapts to local luminance, producing natural, semi-transparent backings.

3.6 Document generation and interactivity

As shown in Fig. 5, our system supports both automatic and interactive use. In automatic mode, $\{s_i\}$ and $\{u_i\}$ drive background generation with masking+ARO. In interactive refinements (e.g., "make the background more subtle"), we recompute u_i and regenerate without modifying text. For structural edits (text content changes), layout boxes are recomputed and the full pipeline reruns with updated masking and ARO.

4 Experiments and Results

4.1 Benchmarking Document Datasets

To evaluate our framework under controlled yet realistic conditions, we constructed a benchmark consisting of academic-style documents and slide decks.

Each instance contains exactly three pages or slides, designed to simulate instructional materials with coherent narrative progression. This structure enables assessment of both per-page visual quality and cross-page stylistic consistency, which is central to background editing in multi-page settings. Each page integrates long-form text, structured bullet points, and at least one embedded figure positioned in non-trivial layouts (e.g., side-anchored, floating, or interleaved with text). These layouts introduce realistic spatial constraints for foreground preservation while avoiding artificial simplifications. By fixing foreground content and layout, the benchmark directly reflects the constrained editing scenario studied in this work. All textual and visual content is authored or generated specifically for this benchmark (text via GPT-5 and images via GPT-4o) to ensure full reproducibility and redistribution. **Constructing a publicly shareable benchmark from real academic papers, textbooks, or presentation slides has limitations, as such materials typically cannot be freely modified or redistributed.** Publicly available "real-world" document corpora are often domain-specific (e.g., government notices or news articles), which would introduce unintended dataset bias. Our synthetic construction therefore enables balanced coverage across layout styles, thematic topics, and page-level narrative structures while maintaining legal clarity and experimental control. Importantly, the benchmark focuses on document editing scenarios where background synthesis is meaningful—namely, materials with intentional margins, structured layouts, and moderate text density typical of lecture slides and academic handouts. We intentionally avoid severely degraded scanned documents or OCR-heavy artifacts, as such settings shift the problem toward document restoration rather than foreground-preserving background generation. This design isolates the core challenges addressed in this paper—layout fidelity, readability preservation, and multi-page stylistic coherence—without conflating them with unrelated noise factors. Additional implementation details and dataset statistics are provided in Appendix A.5.

4.2 Implementation Details

For both PDFs (A4) and slides (16:9), empty dummy images were used as inputs to ensure format alignment, upon which the Summarization Model distilled document text into concise visual themes and the Instruction Generation Model generated editing prompts under multi-page consistency, both model implemented with GPT-4o; bounding boxes were extracted with PyMuPDF and OpenCV, where precise text-region boxes were supplied to ARO for contrast-aware overlays (target contrast 7.0, coverage 0.98, padding 24, radius fraction 0.12), while representative bounding boxes that enclose the overall text regions were provided to Latent Masking to guide attenuation (strength=0.2, start step=0.29 of the diffusion schedule). Unlike ARO, which requires pixel-accurate boxes to guarantee WCAG contrast, Latent Masking only needs representative boxes since its goal is to suppress intrusive background synthesis rather than ensure pixel-level legibility. Latent Masking was implemented inside BAGEL, which serves as a baseline framework for background editing, and our contributions build

Method	Layout ↑	Color ↑	Graphic Style ↑	Compliance ↑	WCAG ↑ (%)	OCR Acc. ↑	CLIP MP Consistency ↑	CLIP Prompt Score ↑	LLM Voting ↑
BAGEL	4.1025	4.275	4.1671	4.325	66.98	0.5536	0.5571	0.1877	4.2292
GPT-5	3.8807	4.0685	4.0050	4.1164	55.02	0.5225	0.6870	0.1687	4.0185
Ours (Prompt+Text mode)	4.2028	4.4285	4.2485	4.4000	**99.75**	**0.9665**	**0.6955**	**0.2374**	**4.3342**
Ours (Prompt only mode)	**4.355**	**4.545**	**4.3478**	**4.7357**	99.38	0.9578	0.6259	0.2042	4.5100
Ours w/o LM	4.2700	4.5407	4.3528	4.7100	99.67	0.9085	0.6905	**0.2542**	4.4907
Ours w/o ARO	4.0835	4.4114	4.2057	4.4992	97.35	0.9012	0.6886	0.2336	4.3214
Ours w/o MPC	4.1664	4.3807	4.2107	4.1985	99.69	0.9632	0.6420	0.2287	4.2592

Table 1: Quantitative evaluation of document background generation across nine metrics. LLM-judged metrics (Layout, Color, Graphic Style, Compliance, LLM Voting) are evaluated on a 1–5 scale by GPT-4o, while automatic metrics (WCAG Contrast Coverage, OCR Accuracy, CLIP MP Consistency, CLIP Prompt Score) measure text readability, accessibility compliance, multi-page consistency, and image–text alignment. Results are reported for BAGEL, GPT-5, and our method in two operating modes (Prompt+Text and Prompt-only), as well as internal ablations. Higher values indicate better performance for all metrics.

upon it without altering the novelty of our readability-preserving and thematically consistent generation pipeline. We note that GPT-4o is used only in the Summarization and Instruction modules, which treat the LLM as a swappable component; LM, ARO, and RNB are LLM-independent. The *Ours w/o MPC* ablation isolates this dependency: removing the LLM-driven module changes only consistency (CLIP MP −0.06, LLM Voting −0.07) while WCAG (99.69%) and OCR (0.96) remain unchanged (Tab. 1).

4.3 Qualitative Results

We evaluate our framework against two representative baselines: BAGEL [13], a document-oriented editing system, and GPT-5 [34], a strong general-purpose multimodal model. Our task differs fundamentally from layout synthesis or poster generation approaches [6, 49], which typically assume flexible structural rearrangement or text re-rendering. In contrast, we address a constrained editing setting in which the **foreground text, figures, and layout of an existing multi-page document must remain unchanged, and only the background layer is modified.** We therefore **restrict comparisons to methods that can operate directly on fixed document pages without regenerating layout elements or retypesetting text.** Including layout-generation pipelines would require architectural adaptation and alter the problem definition, making direct comparison misleading. Figures 3–4 demonstrate that our method preserves structural fidelity and readability while introducing visually coherent backgrounds across styles and page sequences. BAGEL often produces strong textures or high-frequency patterns that partially interfere with dense text regions. GPT-5 can generate aesthetically appealing outputs, but frequently alters existing layout structure, modifies embedded figures, or hallucinates additional text, which violates the foreground-preservation constraint central to our task. Across both Prompt+Text and Prompt-only modes, our approach maintains layout integrity while achieving stylistic consistency across pages. Additional qualitative examples and further discussion on baseline selection are provided in Appendix A.6.

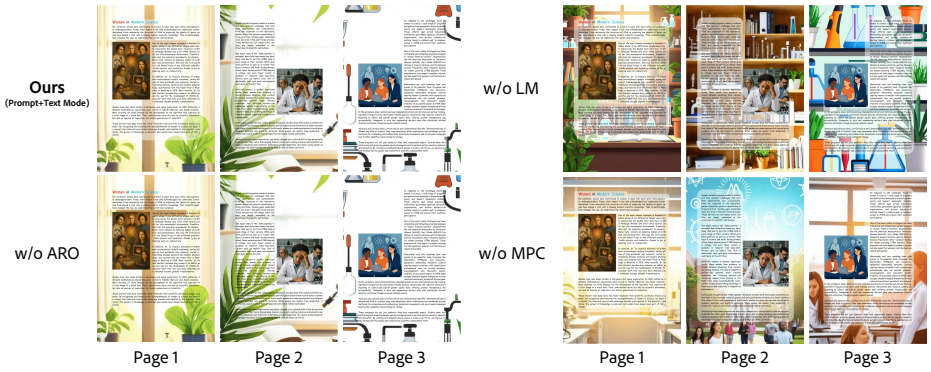

Fig. 6: Ablation on our document-aware background generation. Ours (left) preserves readability and maintains consistent visual themes across pages. **w/o LM** (no latent masking) allows background objects to intrude into foreground text and images; ARO cannot recover readability due to intrusion. **w/o ARO** (no readability optimization) keeps the theme, but text becomes harder to read due to insufficient contrast. **w/o MPC** (no multi-page consistency) produces different styles on each page, losing cross-page visual coherence.

4.4 Quantitative Evaluation

We evaluate against BAGEL and GPT-5 using metrics aligned with prior document-generation work [49]. Our method achieves the best performance across all categories: (i) Design Quality (layout, color, graphic style, compliance), (ii) Readability (WCAG contrast coverage, OCR accuracy), (iii) Multi-page Consistency (CLIP similarity, LLM voting). Our framework reaches near-perfect WCAG compliance (99.75%) and the highest CLIP-based consistency. Detailed analysis are shown in Appendix A.7.

4.5 User Study

We conducted a user study with 30 participants to evaluate our document background generation framework. For each task, participants viewed the original document (PDF or slide) and three anonymized outputs (BAGEL, GPT-5, and

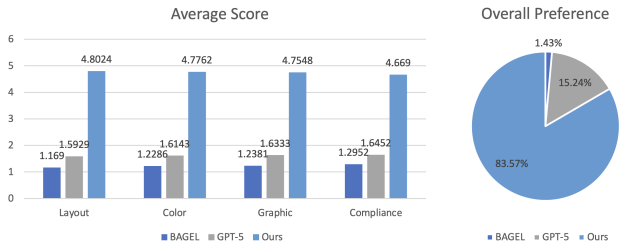

Fig. 7: User study results. Thirty participants evaluated three anonymized systems across four design dimensions: Layout, Color, Graphic Style, and Prompt Compliance (left). Our method achieved the highest score in all categories. In overall preference voting (right), **83.57%** of users selected our result over GPT-5 (15.24%) and BAGEL (1.43%).

Ours) and rated them across four design dimensions—*Layout preservation*, *Color harmony*, *Graphic style consistency*, and *Prompt compliance*—using a 5-point Likert scale. Appropriate measures were taken to ensure that method identities were not disclosed to participants during the study. As shown in Figure 7, our method achieved the highest score in every dimension (4.669–4.8024), whereas BAGEL and GPT-5 scored substantially lower (1.169–1.6452). Participants also selected their preferred output for each task, and **83.57%** of all votes favored *Ours*, compared to 15.24% for GPT-5 and 1.43% for BAGEL. These results demonstrate that users consistently prefer our approach for maintaining readability and producing visually coherent backgrounds. Detailed analysis and explanations are provided in Appendix A.9.

4.6 Ablation Study

We validate the contributions of each component: latent masking (LM), automated readability optimization (ARO), and multi-page consistency (MPC). Removing LM or ARO reduces readability (WCAG and OCR). Removing MPC reduces thematic continuity across pages while preserving readability. Detailed analysis are included in Appendix A.8.

5 Conclusion

In this work, we introduced a multi-layered document editing framework with automated text-conditioned background design to generate visually coherent, text-preserving, and thematically consistent multi-page documents. Our method applies latent masking, which protects text and figures through soft attenuation in the diffusion space, inspired by smooth barrier functions in physics and numerical optimization. This strategy suppresses updates in sensitive regions without applying hard erasure, allowing backgrounds to be regenerated while preserving readability. To further guarantee legibility, we integrate Automated Readability Optimization (ARO), a contrast-driven, WCAG-guided algorithm that adaptively determines the minimal opacity of semi-transparent backings per text region. ARO ensures that inserted shapes remain visually harmonious with the background while meeting accessibility standards, and in combination with latent masking, yields both natural and readable results. Furthermore, we ensure multi-page visual consistency through a recursive summarization-and-instruction process, where each page is distilled into a compact representation that guides subsequent generations, enabling coherent evolution of visual motifs across entire slide decks or reports. Finally, by adopting a layered editing paradigm that treats text, figures, and backgrounds as separate compositional elements, and incorporating prompt-based customization, the framework balances automated coherence with flexible user control, bridging generative modeling and real-world document design workflows. **Limitations and Future Works.** Detailed discussion of limitations and future work is provided in Appendix A.2.

References

1. Ames, A.D., Xu, X., Grizzle, J.W., Tabuada, P.: Control barrier function based quadratic programs for safety critical systems. IEEE Transactions on Automatic Control **62**(8), 3861–3876 (2017). https://doi.org/10.1109/TAC.2016.2638961
2. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18208–18218 (2022)
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
4. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)
5. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM transactions on Graphics (TOG) **42**(4), 1–10 (2023)
6. Chen, H., Xu, X., Li, W., Ren, J., Ye, T., Liu, S., Chen, Y.C., Zhu, L., Wang, X.: Posta: A go-to framework for customized artistic poster generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28694–28704 (2025)
7. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser: Diffusion models as text painters. Advances in Neural Information Processing Systems **36**, 9353–9387 (2023)
8. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser-2: Unleashing the power of language models for text rendering. In: European Conference on Computer Vision. pp. 386–402. Springer (2024)
9. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5343–5353 (January 2024)
10. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
11. Cui, K., Zhang, G., Zhan, F., Huang, J., Lu, S.: Fbc-gan: Diverse and flexible image synthesis via foreground-background composition. arXiv preprint arXiv:2107.03166 (2021)
12. Dalva, Y., Li, Y., Liu, Q., Zhao, N., Zhang, J., Lin, Z., Yanardag, P.: Layerfusion: Harmonized multi-layer text-to-image generation with generative priors. arXiv preprint arXiv:2412.04460 (2024)
13. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
14. Eshratifar, A.E., Soares, J.V., Thadani, K., Mishra, S., Kuznetsov, M., Ku, Y.N., De Juan, P.: Salient object-aware background generation using text-guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7489–7499 (2024)
15. Hauser, J., Saccon, A.: A barrier function method for the optimization of trajectory functionals with constraints. In: Proceedings of the 45th IEEE Conference on Decision and Control. pp. 864–869 (2006). https://doi.org/10.1109/CDC.2006.377331

16. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
17. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
18. Huang, R., Cai, K., Han, J., Liang, X., Pei, R., Lu, G., Xu, S., Zhang, W., Xu, H.: Layerdiff: Exploring text-guided multi-layered composable image synthesis via layer-collaborative diffusion model. In: European Conference on Computer Vision. pp. 144–160. Springer (2024)
19. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., Chen, S.: Diffusion model-based image editing: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(6), 4409–4437 (2025). https://doi.org/10.1109/TPAMI.2025.3541625
20. Inoue, N., Masui, K., Shimoda, W., Yamaguchi, K.: Opencole: Towards reproducible automatic graphic design generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8131–8135 (2024)
21. Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q.: Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In: European Conference on Computer Vision. pp. 150–168. Springer (2024)
22. Kang, T.: Multiple gan inversion for exemplar-based image-to-image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3515–3522 (2021)
23. Kang, T., Lin, M.C.: Action2dialogue: Generating character-centric narratives from scene-level prompts. arXiv preprint arXiv:2505.16819 (2025)
24. Leykin, A., Tuceryan, M.: Automatic determination of text readability over textured backgrounds for augmented reality systems. In: Third IEEE and ACM International Symposium on Mixed and Augmented Reality. pp. 224–230. IEEE (2004)
25. Li, F., Liu, A., Feng, W., Zhu, H., Li, Y., Zhang, Z., Lv, J., Zhu, X., Shen, J., Lin, Z., et al.: Relation-aware diffusion model for controllable poster layout generation. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 1249–1258 (2023)
26. Li, P., Huang, Q., Ding, Y., Li, Z.: Layerdiffusion: Layered controlled image editing with diffusion models. In: SIGGRAPH Asia 2023 Technical Communications, pp. 1–4 (2023)
27. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22511–22521 (2023)
28. Li, Z., Li, F., Feng, W., Zhu, H., Li, Y., Zhang, Z., Lv, J., Shen, J., Lin, Z., Shao, J., et al.: Planning and rendering: Towards product poster generation with diffusion models. arXiv preprint arXiv:2312.08822 (2023)
29. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022)
30. Luo, C., Shen, Y., Zhu, Z., Zheng, Q., Yu, Z., Yao, C.: Layoutllm: Layout instruction tuning with large language models for document understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15630–15640 (2024)
31. Ma, Y., Liu, S., Ma, A., Wu, X., Leng, D., Yin, Y.: Hico: Hierarchical controllable diffusion model for layout-to-image generation. Advances in Neural Information Processing Systems **37**, 128886–128910 (2024)

32. Morita, R., Kuno, S., Tanaka, R., Li, R., Dinh, H.D., Sukeda, I.: Sawna: Space-aware text to image generation. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters. SIGGRAPH Posters '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3721250.3743023, https://doi.org/10.1145/3721250.3743023
33. OpenAI: Gpt-4o: Openai's most advanced generative text and vision model. https://openai.com/index/hello-gpt-4o/ (2024), 2024-05-13
34. OpenAI: Introducing gpt-5. https://openai.com/index/introducing-gpt-5/ (2025), 2025-08-07
35. Peng, Y., Xiao, S., Wu, K., Liao, Q., Chen, B., Lin, K., Huang, D., Li, J., Yuan, Y.: Bizgen: Advancing article-level visual text rendering for infographics generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23615–23624 (2025)
36. Rabiee, P., Hoagg, J.B.: Soft-minimum barrier functions for safety-critical control subject to actuation constraints. 2023 American Control Conference (ACC) pp. 2646–2651 (2023), https://api.semanticscholar.org/CorpusID:257913295
37. Scharff, L.F., Ahumada Jr, A.J.: Contrast measures for predicting text readability. In: Human Vision and Electronic Imaging VIII. vol. 5007, pp. 463–472. SPIE (2003)
38. Scharff, L.F., Hill, A.L., Ahumada Jr, A.J.: Discriminability measures for predicting readability of text on textured backgrounds. Optics express **6**(4), 81–91 (2000)
39. Sun, W., Li, T., Lin, Z., Zhang, J.: Spatial-aware latent initialization for controllable image generation. arXiv preprint arXiv:2401.16157 (2024)
40. W3C World Wide Web Consortium: Web content accessibility guidelines 2.2. W3C Recommendation, 12 Dec 2024 (2024), https://www.w3.org/TR/WCAG22/, confirmed update date via W3C: see "Status of This Document" section
41. Wang, Y., Zhan, S.S., Jiao, R., Wang, Z., Jin, W., Yang, Z., Wang, Z., Huang, C., Zhu, Q.: Enforcing hard constraints with soft barriers: Safe reinforcement learning in unknown stochastic environments. ArXiv **abs/2209.15090** (2022), https://api.semanticscholar.org/CorpusID:252668725
42. Wang, Z., Bao, J., Gu, S., Chen, D., Zhou, W., Li, H.: Designdiffusion: High-quality text-to-design image generation with diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20906–20915 (2025)
43. Weng, H., Huang, D., Qiao, Y., Hu, Z., Lin, C.Y., Zhang, T., Chen, C.: Desigen: A pipeline for controllable design template generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12721–12732 (2024)
44. Wu, K., Chen, J., Liang, Z., Wang, Y., Li, J., Zhang, C., Wang, B., Yuan, Y.: Hybrid layout control for diffusion transformer: Fewer annotations, superior aesthetics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17930–17940 (2025)
45. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7452–7461 (October 2023)
46. Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., Cui, B.: Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In: Forty-first International Conference on Machine Learning (2024)
47. Zhang, H., Hong, D., Yang, M., Cheng, Y., Zhang, Z., Shao, J., Wu, X., Wu, Z., Jiang, Y.G.: Creatidesign: A unified multi-conditional diffusion transformer for creative graphic design. arXiv preprint arXiv:2505.19114 (2025)

48. Zhang, L., Agrawala, M.: Transparent image layer diffusion using latent transparency. arXiv preprint arXiv:2402.17113 (2024)
49. Zhang, Z., Cheng, Y., Hong, D., Yang, M., Shi, G., Ma, L., Zhang, H., Shao, J., Wu, X.: Creatiposter: Towards editable and controllable multi-layer graphic design generation. arXiv preprint arXiv:2506.10890 (2025)
50. Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., Li, X.: Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22490–22499 (2023)
51. Zuffi, S., Brambilla, C., Beretta, G., Scala, P.: Human computer interaction: Legibility and contrast. In: 14th international conference on image analysis and processing (ICIAP 2007). pp. 241–246. IEEE (2007)
52. Zuffi, S., Brambilla, C., Beretta, G.B., Scala, P.: Understanding the readability of colored text by crowd-sourcing on the web. HP Laboratories (2009)