DECOUPLEGEN: INTENT-AWARE COT LAYOUT PLANNING AND CONTROLLABLE TEXT RENDERING FOR VIDEO COVERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video covers are often a viewer's first encounter with a video, yet their typography must be composed over a content-bearing frame whose people, objects, actions, and native screen content remain recognizable. We study this progressive design problem with **DecoupleGen**, which connects frame selection, intent-aware layout planning, and background-conditioned text rendering through an explicit text–quadrilateral–color interface. The frame selector identifies a relevant and visually usable background. The planner first produces a natural-language rationale for title hierarchy, placement, and contrast, then predicts structured text regions and quantized colors. A diffusion renderer converts that plan into rasterized spatial and color conditions and uses localized OCR-feature supervision to strengthen character fidelity. We introduce **Cover-100K**, containing 99,781 aligned examples of recovered clean backgrounds, completed covers, strings, geometry, colors, and generated design rationales. On Cover-600 with supplied clean backgrounds, DecoupleGen improves text accuracy and distributional image quality over released editors, while human raters favor its planned layouts over PosterLlama. Component studies support the contributions of cover-domain supervision and structured rendering. Together, the dataset and structured interface provide a practical path from video content and design intent to readable covers.



Figure 1: DecoupleGen results, with a selected background and its rendered cover in each pair.

1 INTRODUCTION

A video cover is a designed visual summary of a source video. Its production does not begin from an empty canvas or a single prompt, but from a temporal stream containing many moments. The system identifies a frame that is representative of the video and also usable for subsequent design. That choice determines which people, objects, actions, native subtitles, numbers, and screen states must remain recognizable, as well as where new typography can plausibly be placed. Cover generation is therefore a progressive video-to-design problem in which each decision constrains the next.

Existing methods contribute content-aware layout Zheng et al. (2019); Hsu et al. (2023), poster synthesis Lin et al. (2023b); Chen et al. (2025), and layout–rendering separation Chen et al. (2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Yet simply composing them is insufficient: a video cover must preserve its selected frame as source evidence, so a poor frame choice cannot be repaired by rebuilding the scene, and subsequent layout and rendering must remain content-conditioned.

The first decision is therefore *frame acquisition*. A useful frame should be relevant and aesthetically viable, avoid blur and closed eyes, leave important subjects inside the boundary, and retain a strong original-cover candidate. This step converts a temporal video into the evidence-bearing background on which all later design operates.

The second decision is *layout planning*. Once a frame is fixed, a title cannot be placed by geometry alone. The model must interpret the requested text, establish hierarchy across multiple lines, identify regions that do not obscure salient content, and choose colors that remain legible against the local background. We make this reasoning explicit: the planner first states a design rationale about hierarchy, placement, and contrast, and only then expresses that rationale as quadrilateral text regions and discrete color tokens.

The final decision is *controllable rendering*. A correct plan is useful only if the image generator follows its strings, positions, and colors without altering the surrounding frame. We therefore turn the structured layout into spatial and color conditions for diffusion and add localized OCR-feature consistency within the planned text regions. The conditions determine where and how text should appear, while the localized semantic loss concentrates supervision on character identity.

DecoupleGen follows this same progression from video evidence to design intent and then to pixels. Its frame selector proposes a background, its intent-aware planner converts a generated design rationale into an inspectable layout, and its renderer realizes that layout on the background. Cover-100K supplies aligned clean backgrounds, design rationales, structured layouts, colors, and completed covers for supervising the two learned stages. The task diagnosis in Section 3.2 further grounds this formulation by comparing the preservation scope and local text-background conditions of covers with adjacent poster data, while keeping the quantitative analysis separate from the method narrative.

Cover-600 evaluates the learned stages from supplied clean backgrounds using OCR, FID, human preference, and matched ablations. These results measure text generation and perceived design quality under a reproducible prepared-background protocol. Our contributions are:

- We formulate video-cover design around a selected, content-bearing frame and introduce Cover-100K, which aligns clean backgrounds and completed covers with design rationales, text, quadrilateral geometry, and colors. A task diagnosis distinguishes this setting from adjacent poster data.
- We develop intent-aware CoT layout planning that generates a rationale for title hierarchy, placement, and contrast before predicting a structured text–quadrilateral–quantized-color layout.
- We develop controllable diffusion rendering that rasterizes the layout into spatial and color maps and applies localized OCR-feature supervision. Cover-600 evaluates the two learned stages from supplied clean backgrounds.

2 Related Work

Video thumbnail selection. Multi-task visual-semantic embedding Liu et al. (2015), beauty/click prediction Song et al. (2016), sentence-conditioned thumbnail generation Yuan et al. (2019), and multimodal video summarization Qiu et al. (2024) select representative frames from videos. Our inference pipeline combines relevance and quality cues with hard usability vetoes, person- and category-aware adjustments, and an original-cover preference gate.

Content-aware layout generation. ContentGAN Zheng et al. (2019) predicts design layouts. CGL-GAN Zhou et al. (2022) and DS-GAN/PosterLayout Hsu et al. (2023) inject saliency to avoid important content. RADM Li et al. (2023) refines layouts with diffusion. LayoutPrompter Lin et al. (2023a), PosterLlama Seol et al. (2024), and PosterO Hsu & Peng (2025) leverage LLM/MLLM priors. Our Layout MLLM additionally produces a natural-language design rationale and per-element color.



Figure 2: DecoupleGen covers across diverse styles; faces are masked for privacy.

Poster generation. AutoPoster Lin et al. (2023b) predicts layout and visual attributes before rendering text. Prompt2Poster Wang et al. (2024) uses an LLM to plan background and typography. COLE Jia et al. (2023) produces layered SVG designs. PosterMaker Gao et al. (2025), POSTA Chen et al. (2025), and AutoPP Fan et al. (2026) add diffusion rendering, product/background generation, or online optimization. These systems generally design or regenerate the canvas around a product or prompt, whereas a video cover must preserve the semantic evidence already present in its selected frame.

Text-rich generation and image editing. General diffusion and image-editing models can synthesize stylized text from language instructions Esser et al. (2024); Wu et al. (2025); Batifol et al. (2025). TextDiffuser Chen et al. (2023) and TextDiffuser-2 Chen et al. (2024) already combine layout planning and diffusion text rendering; SceneVTG Zhu et al. (2024) plans regions on real backgrounds, and AnyText Tuo et al. (2024) uses localized OCR supervision. Our distinct design is the cover-specific rationale–quadrilateral–quantized-color interface, aligned cover data, and its coupling to a background-conditioned renderer.

3 Video-to-Cover Pipeline and Cover-100K

3.1 Cover-100K Construction

Cover-100K contains 99,781 high-quality samples collected from online video platforms. Its construction starts from completed covers and recovers aligned clean backgrounds I_{bg}, whereas infer-

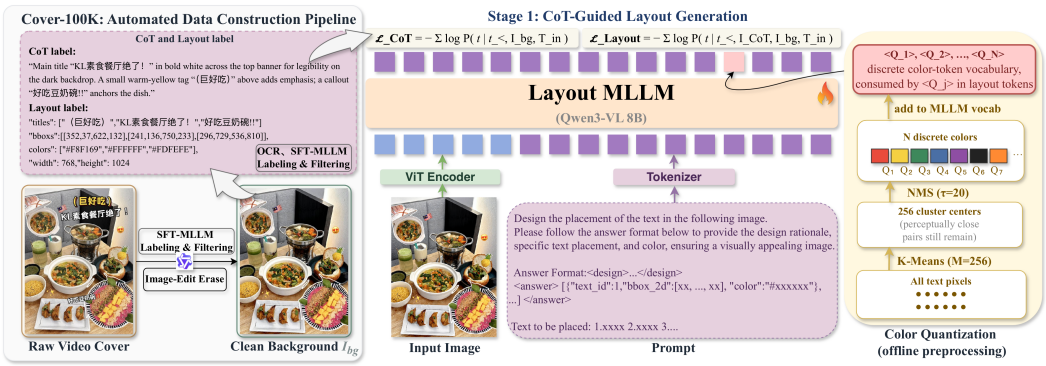

Figure 3: Cover-100K construction and Stage 1 rationale–layout planning with quantized color tokens.

ence selects I_{bg} from a source video. Both paths use the same background–layout interface. The automated construction pipeline has four modules:

- **Data collection and filtering.** Covers are harvested using user-preference signals such as CTR, likes, and playback counts. A binary MLLM filter removes samples with poor layouts, weak visual coherence, or distracting typography.
- **Design-element erasure.** An image-editing model removes text, effects, and underlays to recover a clean background. Unlike regular text masks, real cover effects include irregular strokes, shadows, banners, and translucent underlays; mask-based inpainting can leave these remnants or regenerate characters.
- **Layout annotation.** OCR pre-annotations are manually corrected on a seed subset, including text strings, quadrilateral boxes, and hexadecimal colors. A fine-tuned MLLM then labels the remaining data.
- **CoT annotation.** An MLLM describes hierarchy and spatial/color relations among text, salient content, and background, providing supervision for explicit design reasoning and readable visual composition.

Each record aligns four views: the completed cover I_{cover}, recovered clean background I_{bg}, natural-language rationale I_{CoT}, and structured per-element text, geometry, and color $\mathcal{L}$. Together they form an inter-stage interface, not independent metadata. The rationale and $\mathcal{L}$ supervise Stage 1; I_{bg}, I_{cover}, and rasterized $\mathcal{L}$ supervise Stage 2. At inference, planner output replaces the annotated layout without changing the renderer interface. This lets intent, layout, and pixel synthesis be trained separately while remaining composable.

3.2 Video Cover–Poster Diagnosis

We test whether video covers require a dedicated task, dataset, and pipeline beyond poster generation. Figure 4 contrasts Cover, AutoPP1M Fan et al. (2026), and PosterArt Chen et al. (2025) along preservation-footprint and local-complexity axes.

Preservation footprint. For Cover, non-text regions proxy preserved content; AutoPP1M uses the product alpha mask. Video covers retain most canvas; product posters leave more non-product area for redesign. These dataset-specific proxies characterize the relative editing scope.

Local text-background complexity. We area-weight measurements beneath intended text regions. Compared with PosterArt, Cover has higher entropy and intensity variation but similar edge density, revealing heterogeneous, content-bearing backgrounds rather than simply "more edges." Combined with the larger preservation footprint, this mismatch shows that poster data and pipelines do not fully represent the video-cover setting. It motivates video-derived intent, layout, color, and rendering supervision, frame selection, constrained layout reasoning, and spatial/color-controlled rendering that preserves the frame. These differences motivate dedicated video-cover modeling.

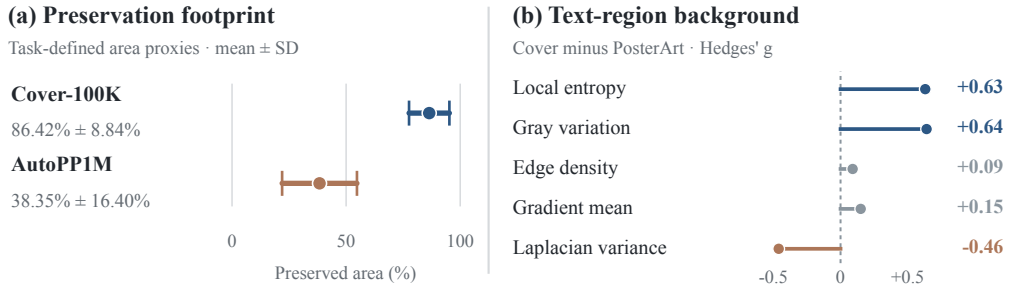

Figure 4: Task diagnosis. (a) Mean preserved area and standard deviation under dataset-specific proxies. (b) Local background statistics at intended text regions, expressed as Hedges' g (Cover relative to PosterArt); positive values indicate higher Cover measurements.

3.3 Inference-time Frame Selection

A video contains many relevant moments, but few usable cover backgrounds. For each candidate, the selector computes clarity, aesthetics, boundary penalty, native-text area, and video relevance. Frames with clarity below 0.45 or aesthetics below 0.30 are rejected. Among valid frames, higher clarity, aesthetics, and relevance increase the score, while boundary violations and excessive native text reduce it.

When a person is detected, face–native-text overlap is penalized and positive or negative expression quality raises or lowers the score; closed eyes reject the frame. Food and pet videos slightly down-weight frames containing people, while no-person frames receive a small compensation in other categories.

The highest-scoring valid frame is selected by default. An original-cover candidate is preferred when it ranks in the top 10%, or when its aesthetics exceeds 0.50 and its clarity exceeds 0.75. The selected frame becomes I_{bg} for intent-aware layout planning and controllable rendering.

4 DecoupleGen

DecoupleGen separates intention, layout, and rendering into two trainable stages. Stage 1 converts the background and request into a rationale and structured layout; Stage 2 renders from geometric and color conditions with an objective of retaining recognizable source content. The separation exposes a controllable inter-stage interface.

Task interface. After frame selection, DecoupleGen receives background I_{bg} and prompt T_{input} containing requested strings and optional style guidance. The goal is to add typography, effects, and underlays without turning I_{bg} into a newly synthesized scene: people, products, actions, and native screen information should remain recognizable. We factor this mapping as

$$\begin{aligned} F_\phi(I_{\text{bg}}, T_{\text{input}}) &= (I_{\text{CoT}}, \mathcal{L}), \\ G_\theta(I_{\text{bg}}, T_{\text{input}}, \mathcal{L}) &= I_{\text{cover}}, \end{aligned} \tag{1}$$

where F_ϕ is the layout model and G_θ the renderer. The explicit interface $\mathcal{L}$ separates variable-length, inspectable design decisions from continuous pixel synthesis. Stage 1 learns from rationale and layout annotations; Stage 2 rasterizes annotated layouts during training and planner outputs at inference. This separation provides stage-specific supervision and exposes text placement and color as explicit intermediate variables.

4.1 Intent-Aware CoT Layout Planning

Given I_{bg} and user prompt T_{input}, Stage 1 translates high-level design intent into two coupled outputs. It first articulates a design rationale and then predicts $\mathcal{L} = \{(T_i, B_i, C_i) \mid i = 1, \ldots, N\}$, where T_i is a requested string, $B_i \in \mathbb{R}^8$ contains four ordered quadrilateral vertices, and C_i is

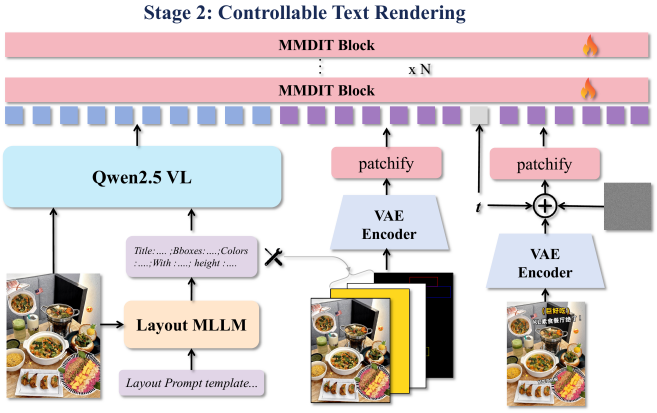

Figure 5: Stage 2 pipeline: layout text and VAE-encoded background, spatial, and color conditions enter MMDiT.

a color token. Strings, quadrilateral geometry, and foreground color are the explicit controls; font family, strokes, banners, underlays, and other effects remain implicit style variables of the renderer. Quadrilaterals retain rotated or perspective-aligned titles that an axis-aligned box cannot express. We employ Qwen3-VL-8B Bai et al. (2025a) as the multimodal backbone, using its visual-language prior to reason jointly about content, free regions, hierarchy, and contrast.

Color quantization. Compression and interpolation make real-world text colors noisy. Directly regressing continuous RGB values would force the autoregressive model to distinguish many visually equivalent perturbations. We extract all training text pixels, run K-means with $M = 256$, and apply RGB-distance non-maximum suppression with threshold $\tau = 20$. Each surviving center becomes a discrete vocabulary token Q_j. An annotated color c_i is assigned by

$$q(c_i) = \arg\min_{Q_j \in \mathcal{Q}} \|c_i - Q_j\|_2, \tag{2}$$

and the predicted token is mapped back to its RGB center when constructing the Stage 2 color map. This makes the autoregressive output robust to small color perturbations and keeps the symbolic and raster interfaces consistent.

Intent-aware CoT generation. Direct layout regression requires the model to emit precise coordinates and colors before making its design logic explicit. We instead supervise an ordered reasoning-and-layout sequence. The model first generates a rationale I_{CoT} describing hierarchy, placement, and contrast, then emits $\mathcal{L}$:

$$\begin{aligned}\mathcal{L}_{\text{Layout}} = &-\log P(I_{\text{CoT}} \mid I_{\text{bg}}, T_{\text{input}}) \\ &-\log P(\mathcal{L} \mid I_{\text{CoT}}, I_{\text{bg}}, T_{\text{input}}).\end{aligned} \tag{3}$$

This text links design intent to geometric and color parameters. Concretely, the target is a two-part sequence: `<design>` I_{CoT} `</design>`, followed by `<answer>[ {"text_id":`i`, "bbox_2d":[...], "color":"<Q_`j`>"}, ... ]</answer>`. The rationale names available regions, salient-object avoidance, title hierarchy, and contrast choices before the machine-parseable records associate every requested T_i with B_i and C_i. The loss in Equation 3 is evaluated on this ordered token sequence, making the rationale an explicit supervised precursor to the layout.

4.2 Controllable Text Rendering

Stage 2 builds on Qwen-Image-Edit to synthesize I_{cover} while following $\mathcal{L}$. In its vanilla form, text can drift from B_i, colors are controlled only coarsely, and characters may be deformed. We address these failures with explicit spatial/color conditioning and localized OCR-feature supervision.

Fine-grained conditional control. We rasterize $\mathcal{L}$ into a binary spatial map M_{spatial} and an RGB color map M_{color}. For pixel x,

$$M_{\text{spatial}}(x) = \mathbb{1}[x \in \bigcup_i B_i],$$
$$M_{\text{color}}(x) = \begin{cases} Q(C_i), & x \in B_i, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{4}$$

where $Q(C_i)$ is the RGB center represented by the predicted token. The clean background and maps are VAE-encoded:

$$z_c = \text{Concat}\big(E(I_{\text{bg}}), E(M_{\text{spatial}}), E(M_{\text{color}})\big), \tag{5}$$

then concatenated with noisy latent z_t and injected into MMDiT. The structured string and user prompt are also encoded by Qwen2.5-VL Bai et al. (2025b). Thus the multimodal tokens carry text identity and style, while the dense latent maps carry geometry and color instead of asking a language prompt to represent precise coordinates. Spatially aligned control maps retain these constraints throughout denoising, including for multi-line or rotated titles. Direct injection also avoids a separate ControlNet-style branch Zhang et al. (2023): the conditions reuse the existing VAE and MMDiT pathway and only expand the conditional input. Spatial and color maps remain separable so that either attribute can be revised without rewriting the other.

Flow and localized OCR supervision. To distinguish noise from velocity, let $z_0 = E(I_{\text{cover}})$, $\eta \sim \mathcal{N}(0, I)$, and $z_t = (1-t)z_0 + t\eta$ for $t \in [0, 1]$. The rectified-flow target is $u_t = \eta - z_0$, and the renderer predicts velocity v_θ:

$$\mathcal{L}_d = \mathbb{E}\|u_t - v_\theta(z_t, c_{\text{te}}, z_c, t)\|_2^2, \qquad \hat{z}_0 = z_t - t v_\theta(z_t, c_{\text{te}}, z_c, t). \tag{6}$$

Here c_{te} is the Qwen2.5-VL text condition, and the decoded estimate is $I'_{\text{cover}} = D(\hat{z}_0)$. The reported CFC-labeled variant additionally compares frozen PP-OCRv3 Li et al. (2022) features from corresponding text quadrilaterals in I_{cover} and I'_{cover}. Appendix B.1 gives a dimensionally valid reference formulation for feature matching; the ablation reports the evaluated CFC-labeled configuration.



Figure 6: Cover-600 comparison: clean background, layout, baselines, DecoupleGen, and reference cover.

5 Experiments

5.1 Setup

Implementation. Layout-MLLM is trained on eight PPU GPUs for five epochs (12 hours), batch size 16, learning rate 10^{-5} with cosine annealing, and maximum side 1024. The renderer is trained on eight A100 GPUs for 20 epochs (10 days) at 1024^2 with learning rate 10^{-4}; $\lambda_c = 0.05$.

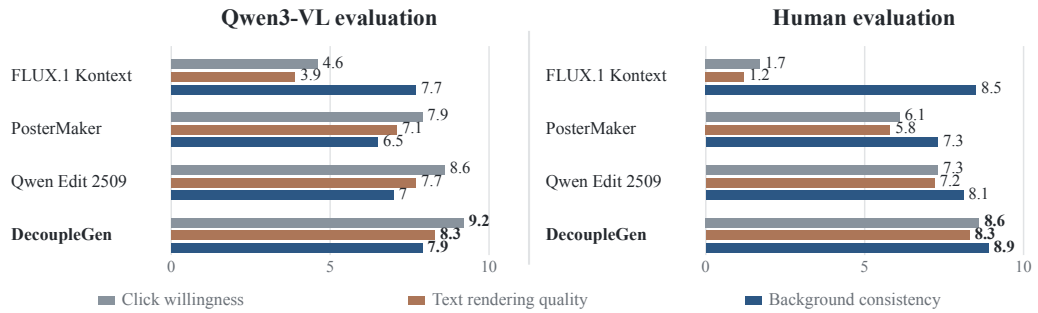

Figure 8: Cover-600 scores for click willingness, text quality, and background consistency from Qwen3-VL (left) and human raters (right).

Evaluation data and metrics. Cover-600 contains 600 high-quality covers disjoint from Cover-100K, spanning background complexities, title lengths, typography styles, and color palettes. The evaluation starts from supplied clean backgrounds and tests the learned planning and rendering stages. Sentence Accuracy (Sen.Acc.) requires an exact OCR match to the requested string, penalizing missing, substituted, or additional characters. Normalized Edit Distance (NED) provides softer string-level similarity, while FID measures distributional image quality.

Subjective protocol. Fifty raters with artistic backgrounds score layout and typography, background consistency, text-rendering quality, and click willingness from 1 to 10. These dimensions distinguish composition, frame preservation, and spelling. We also use Qwen3-VL-235B for pairwise layout diagnosis, treating its scores as auxiliary because it shares a model family with the layout backbone.

Method	FID ↓	Sen.Acc. ↑	NED ↑
FLUX.1 Kontext	57.67	0.1219	0.1967
PosterMaker	51.29	0.7054	0.8144
Qwen-Image-Edit-2509	45.53	0.7821	0.8853
DecoupleGen	**38.41**	**0.8432**	**0.9577**

Table 1: Cover-600 comparison against released checkpoints without Cover-100K fine-tuning. All methods receive the same supplied clean background and text request.

5.2 Baseline Comparisons

We compare released FLUX.1 Kontext Batifol et al. (2025), PosterMaker Gao et al. (2025), and Qwen-Image-Edit-2509 Wu et al. (2025) checkpoints. None is Cover-100K fine-tuned; for every example, all methods receive the same clean background and requested text through their released interfaces. This tests out-of-the-box transfer against released systems, while the matched Qwen-Image-Edit rows in Table 3 separate cover-domain supervision from the added renderer controls.

Objective and qualitative results. Table 1 shows higher OCR scores and lower FID for DecoupleGen in this supplied-background comparison. Figure 6 illustrates spelling, styling, and composition differences.

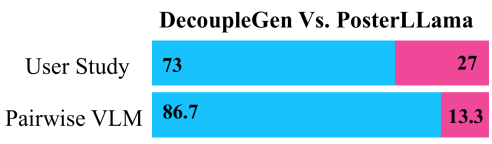

Figure 7: Layout preference on Cover-600.

Layout and preference results. Human raters prefer the predicted layouts to PosterLlama in 73% of pairwise comparisons; Qwen3-VL gives 86.7% (Figure 7). Figure 8 shows higher reported click willingness, text rendering, and perceived background consistency than Qwen-Image-Edit.

5.3 Ablation Study

Comparison	Better	Same	Worse
Qtz vs Base	53.75	5.82	40.43
Qtz+CoT vs Qtz	58.92	2.78	38.30

Table 2: Stage 1 pairwise layout preference (%). Better/Same/Worse rate the first setting against the second; Base uses neither quantized color nor rationale supervision.

Layout planner. Pairwise judgments assess hierarchy, subject avoidance, alignment, and balance. Quantized color is preferred to the base planner in 53.75% of comparisons; adding the rationale sequence is preferred to quantization alone in 58.92% (Table 2). Figure 9 (left) illustrates the resulting color, spacing, and hierarchy; Figure 7 gives the human layout comparison.

Setting	FT	Spa	Col	CFC	FID ↓	Acc. ↑	NED ↑
Qwen Edit	–	–	–	–	45.53	0.7821	0.8853
Qwen Edit	✓	–	–	–	42.75	0.8166	0.9277
+ Spatial	✓	✓	–	–	41.05	0.8202	0.9311
+ Color	✓	–	✓	–	41.72	0.8131	0.9230
+ Spatial + Color	✓	✓	✓	–	39.39	0.8213	0.9301
+ Spatial + Color + CFC	✓	✓	✓	✓	**38.41**	**0.8432**	**0.9577**

Table 3: Stage 2 renderer ablation on Cover-600. FT denotes Cover-100K fine-tuning.

Renderer. Table 3 reports a matched Stage 2 ablation. Cover-100K fine-tuning raises sentence accuracy from 0.7821 to 0.8166 and NED from 0.8853 to 0.9277. Adding spatial and color conditions improves FID to 39.39 while retaining the text gains from cover-domain training. The CFC-labeled configuration further raises sentence accuracy and NED to 0.8432 and 0.9577, respectively. Together, these rows separate the effect of adapting the editor to covers from the effect of providing explicit placement and color controls. Figure 9 (right) connects these aggregate changes to rendered covers, where the planned text regions and colors become visible in the final image.



Figure 9: Qualitative planning (left) and rendering (right) ablations.

6 Conclusion

DecoupleGen connects frame selection, rationale-conditioned layout planning, and diffusion rendering through a structured background–layout interface. The selected frame anchors the visual content; the planner specifies the text, quadrilaterals, and colors; and the renderer turns those conditions into a finished cover. On Cover-600, the resulting covers improve OCR accuracy, and pairwise judgments favor the predicted layouts. The matched ablations attribute complementary gains to cover-domain fine-tuning and explicit rendering conditions. This stage-wise interface also makes the design decisions inspectable: layout judgments address composition, while OCR and image-quality measures characterize readability and overall appearance in the final cover.

AI use statement. Generative AI tools assisted with language polishing, ICLR-template formatting, and clarifying the discussion of existing results. The authors remain responsible.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Haoyu Chen, Xiaojie Xu, Wenbo Li, Jingjing Ren, Tian Ye, Songhua Liu, Ying-Cong Chen, Lei Zhu, and Xinchao Wang. POSTA: A go-to framework for customized artistic poster generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28694–28704, 2025.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. TextDiffuser: Diffusion models as text painters. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *Computer Vision – ECCV 2024*, pp. 386–402. Springer, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Jiahao Fan, Yuxin Qin, Wei Feng, Yanyin Chen, Yaoyu Li, Ao Ma, Yixiu Li, Li Zhuang, Haoyi Bian, Zheng Zhang, Jingjing Lv, Junjie Shen, and Ching Law. AutoPP: Towards automated product poster generation and optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3768–3776, 2026.

Yifan Gao, Zihang Lin, Chuanbin Liu, Min Zhou, Tiezheng Ge, Bo Zheng, and Hongtao Xie. PosterMaker: Towards high-quality product poster generation with accurate text rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8083–8093, 2025.

HsiaoYuan Hsu and Yuxin Peng. PosterO: Structuring layout trees to enable language models in generalized content-aware layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8117–8127, 2025.

HsiaoYuan Hsu, Xiangteng He, Yuxin Peng, Hao Kong, and Qing Zhang. PosterLayout: A new benchmark and approach for content-aware visual-textual presentation layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6018–6026, 2023.

Peidong Jia, Chenxuan Li, Yuhui Yuan, Zeyu Liu, Yichao Shen, Bohan Chen, Xingru Chen, Yinglin Zheng, Dong Chen, Ji Li, et al. COLE: A hierarchical generation framework for multi-layered and editable graphic design. *arXiv preprint arXiv:2311.16974*, 2023.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. PP-OCRv3: More attempts for the improvement of ultra lightweight OCR system. *arXiv preprint arXiv:2206.03001*, 2022.

Fengheng Li, An Liu, Wei Feng, Honghe Zhu, Yaoyu Li, Zheng Zhang, Jingjing Lv, Xin Zhu, Junjie Shen, Zhangang Lin, 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.

Jiawei Lin, Jiaqi Guo, Shizhao Sun, Zijiang Yang, Jian-Guang Lou, and Dongmei Zhang. LayoutPrompter: Awaken the design ability of large language models. *Advances in Neural Information Processing Systems*, 36:43852–43879, 2023a.

Jinpeng Lin, Min Zhou, Ye Ma, Yifan Gao, Chenxi Fei, Yangjian Chen, Zhang Yu, and Tiezheng Ge. AutoPoster: A highly automatic and content-aware design system for advertising poster generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1250–1260, 2023b.

Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3707–3715, 2015.

Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, Bo Li, and Lijuan Wang. MMSum: A dataset for multimodal summarization and thumbnail generation of videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21909–21921, 2024.

Jaejung Seol, Seojun Kim, and Jaejun Yoo. PosterLlama: Bridging design ability of language model to content-aware layout generation. In *European Conference on Computer Vision*, pp. 451–468, 2024.

Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, pp. 659–668, 2016.

Y. Tuo, W. Xiang, J.-Y. He, Y. Geng, and X. Xie. Anytext: Multilingual visual text generation and editing. In *International Conference on Learning Representations*, 2024.

Shaodong Wang, Yunyang Ge, Liuhan Chen, Haiyang Zhou, Qian Wang, Xinhua Cheng, and Li Yuan. Prompt2Poster: Automatically artistic chinese poster creation from prompt only. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10716–10724, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Yitian Yuan, Lin Ma, and Wenwu Zhu. Sentence specified dynamic video thumbnail generation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 2332–2340, 2019.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson W. H. Lau. Content-aware generative modeling of graphic design layouts. *ACM Transactions on Graphics*, 38(4):1–15, 2019.

Min Zhou, Chenchen Xu, Ye Ma, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. Composition-aware graphic layout GAN for visual-textual presentation designs. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 4995–5001, 2022.

Y. Zhu, J. Liu, F. Gao, W. Liu, X. Wang, P. Wang, F. Huang, C. Yao, and Z. Yang. Visual text generation in the wild. In *Computer Vision – ECCV 2024*, pp. 89–106. Springer, 2024.

A Partial-data Pilot

Samples	FID ↓	Sen.Acc. ↑	NED ↑
500	55.65	0.6569	0.7809
5K	54.68	0.6685	0.8022
20K	51.01	0.6943	0.8607
40K	**49.10**	**0.7089**	**0.8899**

Table 4: Partial-data pilot on Cover-600, ending at 40K samples; the series is evaluated under its own training configuration.

B Evaluation Protocol and CFC Specification

Inputs and interface. Cover-600 evaluates planning and rendering from supplied clean backgrounds and requested text; frame selection is the preceding pipeline component described in Section 3.3. Cover-100K backgrounds are recovered from finished covers, while deployment can supply raw selected frames through the same interface. The planner's explicit variables are text identity, quadrilateral vertices, and quantized foreground color; other stylistic effects are rendered implicitly.

Rationale supervision. The CoT target is an automatically generated intermediate rationale. The Stage 1 pairwise comparison measures layout preference for the supervised rationale–layout sequence as a whole.

B.1 Nondegenerate Reference Form of CFC

Let $\mathcal{W}(I, B_p)$ warp the same quadrilateral B_p from target and predicted covers to a common patch canvas, with identical intensity normalization. For frozen PP-OCRv3 level $k \in \{1, 2, 3\}$, let $m_p^{(k)}$ and $\hat{m}_p^{(k)}$ be features of those two patches. Their shapes may vary with k, but agree within each paired level. A dimensionally valid, noncollapsing reference loss is

$$\mathcal{L}^{\star}_{\mathrm{CFC}} = \frac{1}{3P} \sum_{p=1}^{P} \sum_{k=1}^{3} \frac{\|m_p^{(k)} - \hat{m}_p^{(k)}\|_F}{\sqrt{C_k H_k W_k}}. \tag{7}$$

It compares corresponding levels separately with fixed positive weights, avoiding cross-level shape mismatch and zero-weight collapse. Equation 7 is a reproducible reference form supplied independently of the reported checkpoint; Table 3 presents the evaluated CFC-labeled configuration.

Metrics and comparisons. Sentence accuracy measures exact OCR matches, NED measures string similarity, and FID summarizes whole-image distributional quality. The 50-person study supplies perceived-quality scores, while the separate Stage 1 pairwise comparison diagnoses layout preference. The released-system table uses off-the-shelf checkpoints; the within-Qwen ablation separates cover-domain fine-tuning from added renderer conditions. Table 4 is a separate partial-data pilot ending at 40K examples and is interpreted within that series.

C Additional Cover–Poster Analysis

The task-diagnosis axes are illustrated in Figure 4. This section reports the corresponding protocols and distributional statistics. AutoPP1M is the product-poster dataset released with AutoPP Fan et al. (2026); PosterArt is the poster dataset introduced with POSTA Chen et al. (2025). All sample counts below are the valid images used in the corresponding analysis, rather than the full dataset sizes.

C.1 Preservation-footprint Distribution

For Cover, the complement of the union of annotated text quadrilaterals is the preservation proxy. For AutoPP1M, the proxy is the region with alpha value greater than 1 in the supplied product image. Table 5 reports the full distributional summary used for the comparison.

Statistic	Cover	AutoPP1M
Valid images	10,000	9,989
Mean $\pm$ std.	$86.42\% \pm 8.84\%$	$38.35\% \pm 16.40\%$
25th percentile	82.50%	26.37%
Median	88.33%	36.87%
75th percentile	92.60%	48.75%
Mean editable/generated	13.58%	61.65%
Hedges' g	3.65	

Table 5: Preservation-footprint distributions. The proxies describe task-defined editing scope and are not identical segmentation annotations.

Metric	Cover	PosterArt	Δ	g
Edge density	0.0377 ± 0.0435	0.0339 ± 0.0512	+11.3%	0.09
Local entropy	6.0167 ± 1.4826	5.0779 ± 1.7143	+18.5%	0.63
Gradient mean	0.1256 ± 0.0967	0.1112 ± 0.0987	+13.0%	0.15
Gray standard deviation	0.1493 ± 0.0756	0.1011 ± 0.0707	+47.7%	0.64
Laplacian variance	241.17 ± 488.56	474.36 ± 959.83	-49.2%	-0.46

Table 6: Area-weighted local statistics at intended text locations (mean $\pm$ standard deviation).

The mean difference is therefore also present in the medians and central quartiles. The Cover proxy does not include underlays, stickers, or decorations outside the text quadrilaterals and consequently overestimates preserved area. To reduce its mean preserved area to the AutoPP1M mean, however, these omitted elements would have to occupy an additional 48.07 percentage points of the canvas on average. The comparison therefore remains substantial under a wider accounting of cover design elements.

C.2 Local Text-background Statistics

For Cover, measurements are computed on the source video frame beneath each ground-truth text quadrilateral; for PosterArt, they are computed on the provided text-free background at the final text locations. All images are resized to a maximum side of 1024 pixels, each region is expanded by 5%, and box measurements are area-weighted within each image. The analysis contains 9,997 Cover images with 27,569 text boxes and 353 PosterArt images with 990 boxes. The five measurements are Canny edge-pixel density, grayscale entropy, mean Sobel-gradient magnitude, grayscale standard deviation, and Laplacian-response variance.

The entropy quartiles (25th, 50th, and 75th) are 5.50, 6.50, and 7.03 for Cover, compared with 4.23, 5.56, and 6.34 for PosterArt. The corresponding gray-standard-deviation quartiles are 0.095, 0.154, and 0.204, compared with 0.048, 0.090, and 0.144. Edge density and gradient mean have smaller standardized differences, while PosterArt has higher Laplacian variance. Together with the entropy and intensity results, these measurements identify distinct local background conditions at intended text locations and complement the preservation-footprint comparison.

D Additional Subjective-evaluation Details

For the score-based study summarized in the main paper, outputs generated from the same background were shown side by side. The 50 raters had artistic backgrounds and received written definitions for each criterion and the low and high ends of the 1–10 scale before evaluation. Raters could pause or withdraw at any time, viewed only publicly available video frames with overlaid design elements, and provided no personally identifying information. Compensation was set above the applicable local minimum wage. The layout comparison was conducted separately as a pairwise choice between the Layout MLLM and PosterLlama.