ReContraster: Making Your Posters Stand Out with Regional Contrast

Peixuan Zhang[#1], **Zijian Jia**[#1], **Ziqi Cai**[2,3], **Shuchen Weng**[*4,2], **Si Li**[*1,5], **Boxin Shi**[2,3]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications

[2]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

[3]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[4]Beijing Academy of Artificial Intelligence

[5]Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance

{pxzhang, jiazijian, lisi}@bupt.edu.cn, zqtsai@gmail.com, {shuchenweng, shiboxin}@pku.edu.cn

Abstract

Effective poster design requires rapidly capturing attention and clearly conveying messages. Inspired by the "contrast effects" principle, we propose ReContraster, the first training-free model to leverage regional contrast to make posters stand out. By emulating the cognitive behaviors of a poster designer, ReContraster introduces the compositional multi-agent system to identify elements, organize layout, and evaluate generated poster candidates. To further ensure harmonious transitions across region boundaries, ReContraster integrates the hybrid denoising strategy during the diffusion process. We additionally contribute a new benchmark dataset for comprehensive evaluation. Seven quantitative metrics and four user studies confirm its superiority over relevant state-of-the-art methods, producing visually striking and aesthetically appealing posters.

1 Introduction

"It is in the contrast of light and dark that design happens."

–Helen Van Wyk

Posters serve as a common medium for achieving specific communicative goals (MacIntosh-Murray, 2007) (*e.g.*, advertising, event promotion, and public campaigns). Unlike art forms that encourage prolonged reflection (*e.g.*, sculptures, poetry, or documentaries), posters must convey their message and capture attention almost instantly (Utoyo et al., 2021). Recent text-to-poster models made significant strides in producing aesthetically coherent layouts (Inoue et al., 2024; Lin et al., 2024; Yang et al., 2024) and detailed textual elements (Chen et al., 2024a; Tuo et al., 2024; Ma et al., 2025). However, significant challenges remain in generating posters that are both visually compelling and communicatively effective from a user's perspective.

The principle of "contrast effects" (Palmer and Gore, 2014; Scherer and Lambert, 2009; O'Connor, 2015) suggests that the brain responds strongly to contrasting stimuli, quickly processing significant differences between objects and triggering heightened neural activation. David Ambarzumjan's artwork Recover (from the "Brushstrokes in Time" series) exemplifies this principle, where the contrast between dark cityscapes and vibrant nature invites viewers to explore the interplay of the depicted elements[1], amplifying both visual impact and thematic clarity.

Applying the "contrast effect" is a promising approach for creating striking posters. However, existing methods (Esser et al., 2024; Inoue et al., 2024; Yang et al., 2024) struggle to reason about contrastive elements from the user-provided theme while preserving visual harmony, often producing disjointed results. In this paper, we propose **ReContraster** (**Re**gional **Contrast** Poste**r**) to address this problem. As illustrated in Fig. 1, ReContraster has the following primary properties: *(i)* **Element contrast:** ReContraster crafts visually striking results by designing contrastive elements and colors (Fig. 1 (a), contrast between ancient towns and modern cities). *(ii)* **Aesthetic harmony:** ReContraster produces aesthetically appealing posters with well-balanced and organized layouts (Fig. 1 (b), shared silhouette between a skull and an island). *(iii)* **Boundary coherence:** ReContraster synthesizes posters with smooth and artifact-free regional transitions based on user-specified region divisions (Fig. 1 (c), transition from a greenery landscape to a burned one).

To the best of our knowledge, ReContraster is the first training-free model capable of generating posters with all the properties above. By emulating the cognitive behaviors of a poster designer, ReContraster introduces the **compositional multi-**

[#] Equal contributions. * Corresponding authors.

[1]We present representative artworks in Sec. A.

Figure 1: Illustration of our ReContraster for poster generation. Given a text description of the theme and visual texts, and a mask indicating region divisions, our model produces: (a) Visually striking poster designs with highly contrastive elements, compared to PosterGen. (b) Structurally balanced compositions with aesthetic harmony, compared to OpenCOLE. (c) Seamless and coherent transitions across region boundaries, compared to SD3.

agent system, which comprises three specialized agents: *(i)* A cognition agent establishes elemental contrast by reasoning about the theme to identify contrastive element pairs; *(ii)* An arranger agent organizes the layout to ensure aesthetic harmony among the selected contrastive element pairs; and *(iii)* A refiner agent iteratively optimizes the poster, aiming to strike a balance between compelling visual contrast and overall compositional harmony. To further address the challenge of boundary coherence, we introduce the **hybrid denoising strategy**. During the initial denoising steps, a gradient consistency loss is applied to penalize gradient differences across boundaries. Subsequently, we employ a joint regional denoising that linearly blends noise from adjacent regions to guarantee harmonious transitions.

For comprehensively evaluating ReContraster, we contribute a new benchmark dataset containing diverse images, each with a corresponding text description and a region division annotation. We conduct four human evaluation experiments to ensure element contrast, content continuity, division accuracy, and text-image consistency. Our contribution could be summarized as follows:

- We propose ReContraster, the first training-free model to leverage regional contrast to make posters stand out. Extensive experiments demonstrate superior performance.
- We develop compositional multi-agent system that constructs three agents to separately identify elements, organize layout, and evaluate generated poster candidates.
- We design hybrid regional denoising to maintain boundary coherence across region boundaries. A benchmark dataset is contributed for comprehensive evaluation.

2 Related Work

2.1 Poster Generation

Poster generation aims to craft visual impact with message clarity. Existing methods can be divided into two main categories: content-aware and complete poster generation. On the one hand, content-aware methods (Zhou et al., 2022; Zheng et al., 2019; Hsu et al., 2023) focus on arranging elements on a given background. This approach has been recently advanced by LLM-based models (*e.g.*, LayoutPrompter (Lin et al., 2024) and PosterLLaVa (Yang et al., 2024)), which provide a more nuanced semantic understanding. However, the reliance on a pre-selected background fundamentally limits their creative flexibility and application scenarios. On the other hand, complete poster generation methods offer a more versatile solution, creating posters solely from text prompts (*e.g.*, PosterGen (Yang et al., 2024), COLE (Jia et al., 2023), and OpenCOLE (Inoue et al., 2024)). Although this removes the constraint of a background image, they often struggle to generate aesthetically compelling posters. To overcome these challenges, our work focuses on complete poster generation, with the additional objective of enhancing visual impact through regional contrast.

2.2 Large Language Model Agent

Large Language Models (LLMs) have evolved into powerful agents capable of leveraging reasoning and planning abilities for a wide range of

tasks (Wang et al., 2025b), including basic human activities (*e.g.*, searching (Yao et al., 2023; Qiao et al., 2024; Zhang et al., 2026c), communication (Deng et al., 2024; Zhang et al., 2025d), and visual processing (Achiam et al., 2023; Liu et al., 2023; Cheng et al., 2024)) and more complex skills (*e.g.*, logical and mathematical reasoning (Wei et al., 2022; Zhang et al., 2026b, 2025e; Wang et al., 2025a), task decomposition (Zhang et al., 2024, 2025b; Long et al., 2026), and tool utilization (Wang et al., 2024a; Sun et al., 2026)). With these advancements, LLM agents are introduced into poster design for the generation of visual texts (*e.g.*, COLE (Jia et al., 2023) and OpenCOLE (Inoue et al., 2024)) and poster layout (*e.g.*, LayoutPrompter (Lin et al., 2024), PosterLlama (Seol et al., 2024), PosterLLaVa (Yang et al., 2024), and postermaker (Gao et al., 2025)). However, existing methods are primarily single-agent systems that struggle with the multifaceted nature of poster design, particularly in balancing visually striking elements with aesthetically appealing compositions. To address this limitation, we propose a multi-agent system to collaboratively identify contrastive elements, organize poster layout, and achieve a balance between regional contrast and overall harmony.

2.3 Text-driven Image Generation

Text-driven image generation has become a central focus of recent visual content synthesis (Ding et al., 2021; Ramesh et al., 2021; Weng et al., 2023). Recent progress has been significantly accelerated by diffusion models, with notable examples like Stable Diffusion (Rombach et al., 2022; Podell et al., 2023; Chen et al., 2024b; Zhang et al., 2025c, 2026a) generating high-quality images by iteratively denoising Gaussian noise conditioned on textual prompts. TextDiffuser-2 (Chen et al., 2024a) and GlyphDraw (Ma et al., 2025) enhance visual text clarity, while AnyText (Tuo et al., 2024) and UDiffText (Zhao and Lian, 2024) progressively expand support to multiple languages. To further improve the controllability, researchers explore the use of adapters to tailor diffusion models for specific tasks (*e.g.*, attribute control (Mou et al., 2024; Ye et al., 2023), style control (Wang et al., 2024b), and spatial control (Zhang et al., 2025a)). Concurrently, alternative approaches (Chang et al., 2023; Voynov et al., 2023) focus on refining the sampling strategy to improve image quality and consistency. To generate regional contrast posters with boundary coherence, we follow prior research to propose the hybrid regional denoising, a novel sampling inference to produce visually harmonious transitions across region boundaries.

3 Methodology

3.1 Overview

As shown in Fig 2, ReContraster is systematically organized into the following stages:

(i) **User input and information extraction.** Initially, users provide a text description and a corresponding mask indicating region divisions. We use an LLM (Achiam et al., 2023) to extract the theme and visual texts separately.

(ii) **Element identification and layout organization.** This stage is handled by our compositional multi-agent system through an iterative process. Specifically, the extracted theme is passed to the cognition agent, which identifies potential contrastive element pairs. Subsequently, the arranger agent selects appropriate element pairs and organizes them into a conceptual poster layout. This proposed layout then guides an initial image generation in *Stage (iii)*. The resulting image is fed back to the refiner agent to evaluate its overall contrast and harmony. If the result is unsatisfactory, the refiner provides a feedback to the cognition and arranger agents to generate an improved results. This loop continues until the refiner agent approves the composition.

(iii) **Image generation and hybrid denoising.** Guided by the approved layout from *Stage (ii)*, a pre-trained diffusion model is used to concrete the selected visual elements (Eq.(1)). To ensure seamless regional transitions, we then apply our hybrid denoising strategy. This involves using a gradient consistency loss (Eqs. (2-3)) for content continuity and a joint regional denoising technique for harmonious transitions (Eqs. (4–7)).

(iv) **Text rendering and final output.** In the final stage, the visual texts specified in *Stage (i)* are rendered onto the generated image by adopting the text rendering method (*i.e.*, OpenCole (Inoue et al., 2024)), producing a visually striking and aesthetically appealing poster.

3.2 Compositional Multi-agent System

To emulate the cognitive behaviors of a poster designer, we propose the compositional multi-agent system, as shown in Fig. 2 (a). Specifically, we introduce three specialized agents: cognition agent,

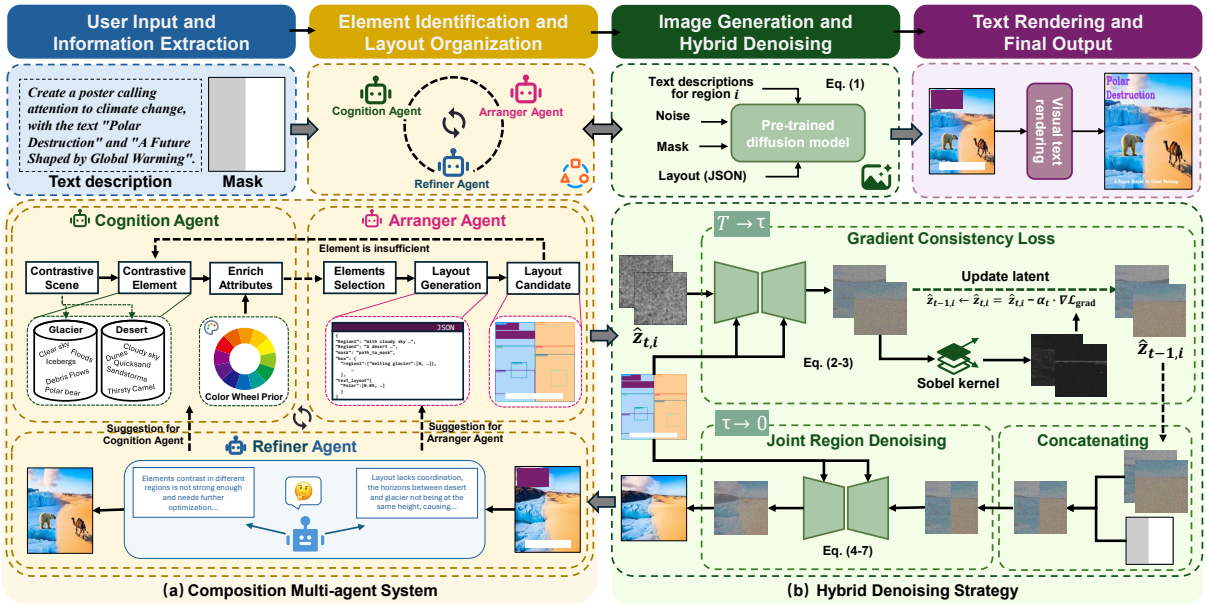

Figure 2: Given a text description and a mask indicating region divisions, ReContraster initially uses an LLM (Achiam et al., 2023) to extract the theme and visual texts separately. The extracted theme is then processed by the **compositional multi-agent system** in an iterative refinement loop. Within each iteration, the cognition agent identifies contrastive elements, the arranger agent organizes a layout in a JSON format, and a candidate image is generated based on this layout. This candidate is then evaluated by the refiner agent, which provides feedback to guide subsequent rounds of improvement until a design is approved. The generation of these candidate images relies on a pre-trained diffusion model equipped with the **hybrid denoising strategy**, where the gradient consistency loss and joint region denoising are applied to maintain content continuity and ensure harmonious transitions across region boundaries, respectively. The visual texts are finally rendered onto the generated image to produce a visually striking and aesthetically appealing poster.

arranger agent, and refiner agent, which collaborate in an iterative loop to design the poster's visual content and layout.

Cognition agent. The cognition agent is responsible for transforming the extracted theme into a set of visually contrastive elements. It utilizes an LLM (Achiam et al., 2023) through a Chain-of-Thought (CoT) (Kojima et al., 2022; Li et al., 2024) process. Specifically, for a given theme T (*e.g.*, calling attention to climate change), the LLM first identifies a pair of high-level contrastive scenes (*e.g.*, glacier and desert) to establish the core thematic conflict. Based on these scenes, the LLM then brainstorms specific objects and forms initial contrastive element pairs (*e.g.*, sandstorm and icebergs). Finally, the LLM enriches these element pairs with detailed color attributes, drawing upon principles from the color wheel (Cohen-Or et al., 2006) are incorporated as prior knowledge. The color wheel defines color relationships by position, distinguishing between complementary colors (*e.g.*, orange and blue) and analogous colors (*e.g.*, purple and blue). The LLM is instructed to apply these principles by assigning complementary colors to the core elements to maximize their visual opposition (*e.g.*, a yellow sandstorm and a blue iceberg), while leveraging analogous colors to ensure harmonious transitions at region boundaries.

Arranger agent. The arranger agent selects optimal element pairs from the candidates provided by the cognition agent and organizes them into an aesthetically harmonious composition. The final layout is structured as a JSON format, generated by simultaneously optimizing for several critical factors, including *shape harmony* by considering element shapes for smooth boundary transitions; *style unification* by ensuring consistent visual styles, textures, and color schemes; *semantic clustering* by grouping related elements for narrative logic; and *typographic integration* by allocating sufficient space for effective text placement. Following these principles, the arranger agent produces an aesthetically harmonious layout to guide the image generation process. Furthermore, when the initial elements prove insufficient to meet these compositional criteria, the agent can issue a targeted request to the cognition agent for additional element pairs.

Refiner agent. The refiner agent serves as a crit-

ical evaluator, guiding the iterative refinement of the generated image. Upon receiving an image, the agent assesses two key aspects: *Visual contrast.* The agent evaluates the clarity and distinctness of elements. If it detects insufficient contrast, it instructs the cognition agent to regenerate element pairs with greater visual differentiation. *Aesthetic harmony.* The agent examines the overall composition for aesthetic imperfections. If improper alignment or imbalanced visual weight are identified, it provides the arranger agent with corrective guidance for optimization. This iterative feedback loop continues until the image satisfies all predefined quality thresholds or a maximum number of iterations is reached.

3.3 Hybrid Denoising Strategy

To concrete contrastive elements in their corresponding regions within the poster, we propose the two-stage hybrid denoising strategy, as illustrated in Fig. 2 (b). Given a pre-trained diffusion model, the first stage focuses on concreting visual content based on the selected and arranged elements, while a gradient consistency loss ensures content continuity across region boundaries. The second stage performs joint denoising across all regions, achieving harmonious transitions and overall visual coherence for the poster.

Gradient consistency loss. To ensure that visual content corresponds accurately to the contrastive elements, we independently concrete visual content during the initial τ denoising steps. Given the latent code $\hat{z}_{t,i}$ at denoising step t and the corresponding text descriptions $\mathcal{P}_i$ of the visual content at region i, the denoising process is formulated as:

$$\hat{z}_{t-\Delta t,i} = \hat{z}_{t,i} - \Delta t \cdot v_\theta(\hat{z}_{t,i}, t, \mathcal{P}_i), \tag{1}$$

where v_θ represents the pre-trained velocity prediction model (Esser et al., 2024) parameterized by θ, t denotes the continuous time in the ODE framework, and Δt is the integration step size for the solver.

However, simply concatenating separately denoised latent codes often leads to visible discontinuities at region boundaries. Therefore, we introduce a gradient consistency loss applied after each denoising step to enforce harmonious transitions by penalizing gradient differences across these boundaries. Specifically, we calculate gradient matrices using a 5×5 Sobel kernel (Levkine, 2012) along boundaries according to region divisions. The gradient consistency loss is defined as:

$$\mathcal{L}_{\text{grad}} = \sum_{b \in \mathcal{B}} \Big(1 - \text{Avg}\big(\text{Cos}(\mathcal{G}_{t,b}, \mathcal{G}'_{t,b})^2\big)\Big), \tag{2}$$

where $\mathcal{B}$ is the set of boundary indices, $\mathcal{G}_{t,b}$ and $\mathcal{G}'_{t,b}$ denote the calculated gradient matrices at denoising step t for the regions along boundary index b, Cos represents the cosine similarity, and Avg denotes the averaging operator. This loss encourages gradient alignment across boundaries. We update each latent code as follows:

$$\hat{z}_{t,i} = \hat{z}_{t,i} - \alpha_t \cdot \nabla \mathcal{L}_{\text{grad}}. \tag{3}$$

After the initial τ steps, the latent codes are concatenated based on the region divisions. We empirically set $\tau = 10$.

Joint region denoising. While the gradient consistency loss ensures content continuity, it does not guarantee harmonious transitions and visual coherence. Considering that the first stage (the initial τ steps) provides a coarse layout of the visual elements, the second stage focuses on the overall harmonization of the different regions to improve the harmonization of the generated poster.

Denoting the binary mask for the region i as $\mathcal{M}_i$, we first concatenate the denoised latent codes at denoising step τ based on the region divisions:

$$z_\tau = \sum_{i \in \mathcal{V}} (\hat{z}_{\tau,i} \odot \mathcal{M}_i), \tag{4}$$

where $\mathcal{V}$ is the set of regions and $\odot$ is element-wise multiplication. Subsequently, we require the visual content in regions near the boundaries to blend smoothly. Let r represent the boundary margin and $d_{i,j}$ represent the shortest distance from each point in region i to the boundary of the adjacent region j, we calculate the distance-weighted noise predictions:

$$\hat{\epsilon}_{i,j} = \frac{r + d_{i,j}}{2r} \cdot v_\theta(z_t, t, \mathcal{P}_i) + \frac{r - d_{i,j}}{2r} \cdot v_\theta(z_t, t, \mathcal{P}_j), \tag{5}$$

where $\mathcal{P}_i$ and $\mathcal{P}_j$ are the text descriptions for region i and j, respectively, and $d_{i,j}$ is clipped to the range $[0, r]$. Next, we incorporate the influence of all adjacent regions following the classifier-free guidance calculation (Ho and Salimans, 2022):

$$\hat{\epsilon}_i = v_\theta(z_t, t, \emptyset) + \frac{w}{|\mathcal{A}(i)|} \cdot \sum_{j \in \mathcal{A}(i)} \big(\hat{\epsilon}_{i,j} - v_\theta(z_t, t, \emptyset)\big), \tag{6}$$

where $\mathcal{A}(i)$ is the set of regions adjacent to region i. Finally, we reformulate the denoising process as:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(z_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \cdot \sum_{i \in \mathcal{V}} (\hat{\epsilon}_i \odot \mathcal{M}_i)\big) + \sigma_t \epsilon. \tag{7}$$

In practice, we set $r = 1/32$ of the latent space size and $w = 3$ according to experimental experience.

4 Benchmark Dataset

Existing datasets for poster generation provide annotations for poster structure layout (Yang et al., 2024; Zhou et al., 2022), visual text arrangements (Lin et al., 2023), and theme descriptions (Xu et al., 2024). However, they primarily focus on movie posters and human figures, generally lacking regional contrast, a critical feature for crafting visually striking posters. To address this limitation, we collect a high-quality benchmark dataset tailored for regional contrast posters. Specifically, we carefully select samples from existing datasets (Yang et al., 2024; Zhou et al., 2022; Lin et al., 2023; Xu et al., 2024), e-commerce platforms, and Instagram, ensuring all posters are used in accordance with their respective terms of service and licensing agreements. We utilize GPT-4o (Achiam et al., 2023) and SAM (Kirillov et al., 2023) to annotate text descriptions and region divisions for each poster, followed by manual verification and refinement. This process results in a dataset of 643 posters for model evaluation[2].

We conduct four human evaluation experiments to assess the quality of our benchmark dataset, focusing on whether: the collected posters contain sufficient element contrast (Exp-I); the collected posters demonstrate the overall aesthetic harmony (Exp-II); the annotated text descriptions accurately describe the theme of each poster (Exp-III); and the annotated region divisions accurately reflect the structure of each poster (Exp-IV). For each experiment, we randomly select 100 samples from our dataset and ask 25 volunteers to independently evaluate these samples, assigning each a rating of "Failed", "Borderline", "Acceptable", or "Perfect". As shown in Tab. 1, over 90% of volunteers rate the annotation quality as "Acceptable" or higher across all four aspects, confirming the dataset quality.

Table 1: Percentage (%) of user ratings in the four experiments of human evaluation for the collected dataset.

Rating	Exp-I	Exp-II	Exp-III	Exp-IV
Failed	1.28	0.44	0.00	0.84
Borderline	6.92	6.44	5.68	2.60
Acceptable	36.56	29.64	25.72	10.92
Perfect	**55.24**	**63.48**	**68.60**	**85.64**

5 Experiments

5.1 Experimental Setup

Since ReContraster is a training-free model[3], it is capable of seamlessly integrating advanced models for its pipeline. In our experiments, we use GPT-4o (Achiam et al., 2023) as the large language model, CreatiLayout (Zhang et al., 2025a) for image generation, and OpenCOLE (Inoue et al., 2024) for text rendering[4]. Notably, since related models are not designed for regional contrast, we prompt the GPT-4o (Achiam et al., 2023) to directly describe the selected elements in each region and their corresponding visual effects. This setup potentially favors the comparison methods.

5.2 Quantitative Evaluation Metrics

We comprehensively evaluate ReContraster across following aspects: *(i)* Following LAION-5B (Schuhmann et al., 2022), we calculate the **LAION Aesthetic Score (LAS)** to measure the aesthetic quality of generated posters. *(ii)* Following StyTr2 (Deng et al., 2022), we compute the **Regional Style Difference (RSD)** by extracting style features across regions using the VGG model (Simonyan and Zisserman, 2015) and calculate the MSE, measuring the visual contrast. *(iii)* Using the Sobel kernel (Levkine, 2012), we calculate the **Boundary Gradient Difference (BGD)** by calculating the cosine similarity between gradients of patches along the region boundaries, measuring the content continuity. *(iv)* Following TextDiffuser (Chen et al., 2023), we recognize visual texts on the poster and compute the **Optical Character Recognition (OCR)** accuracy compared to the provided text, measuring the accuracy of text rendering. *(v)* To further evaluate the holistic quality of the generated posters, we employ the VLM (Achiam et al., 2023) to rate **Content Relevance and Effectiveness (CRE)**, **Visual Appeal and Impact (VAI)**, and **Boundary Integration and Harmony (BIH)** on a scale of 1 to 10 score.[5]

5.3 Comparison with Relevant Methods

As the first approach to regional contrast poster generation, we conduct comparison experiments with related text-to-image generation methods (*i.e.*, SD3 (Esser et al., 2024), Flux.1-dev (Labs, 2024), and TextDiffuser-2 (Chen et al., 2024a)) and complete poster generation methods (*i.e.*, OpenCOLE

[2]The benchmark dataset will be released upon publication, with representative samples visualized in Sec. I.

[3]The analysis of model efficiency is detailed in Sec. D.

[4]The workflow and prompts are provided in Sec. K.

[5]Details are provided in Sec. F.

Figure 3: Visual quality comparisons with text-to-image generation methods and poster generation methods.

Table 2: Quantitative experiment results of comparison, ablation, and user study. ↑ (↓) means higher (lower) is better. The best performances are highlighted in **bold**.

Method	Quantitative Evaluation Metrics							User Study			
	LAS ↑	RSD ↑	BGD ↓	OCR ↑	CRE ↑	VAI ↑	BIH ↑	Exp-I	Exp-II	Exp-III	Exp-IV
SD3	4.2810	596.69	0.0566	0.55	7.61	7.42	6.30	11.76	14.80	15.16	2.04
Flux.1-dev	4.8938	602.49	0.0460	0.58	7.41	7.64	6.62	12.36	16.48	12.24	3.08
TextDiffuser-2	4.3693	446.47	0.0818	0.56	5.30	6.76	4.85	5.04	5.32	3.12	0.92
OpenCOLE	3.6768	286.74	0.0548	0.57	7.45	5.82	6.33	3.28	7.24	14.20	1.48
PosterGen	3.4895	717.80	0.0674	0.43	7.52	5.11	3.17	14.12	10.84	13.08	1.92
PosterMaker	4.5686	600.94	0.0759	0.53	7.02	6.69	4.26	12.08	5.68	10.96	1.20
ReContraster	**5.0966**	**842.60**	**0.0375**	**0.65**	**7.82**	**7.87**	**7.04**	**41.36**	**39.64**	**31.24**	**89.36**
W/o CAI	4.7501	766.02	0.0388	0.56	7.04	7.64	6.94	*N/A*	*N/A*	*N/A*	*N/A*
W/o IRA	4.7280	676.95	0.0446	0.57	7.33	7.57	6.73	*N/A*	*N/A*	*N/A*	*N/A*
W/o GCL	4.9363	789.86	0.0482	0.60	7.38	7.71	6.64	*N/A*	*N/A*	*N/A*	*N/A*
W/o JRD	4.8387	804.52	0.0467	0.62	7.51	7.68	6.58	*N/A*	*N/A*	*N/A*	*N/A*

(Inoue et al., 2024), PosterGen (Yang et al., 2024), and PosterMaker (Gao et al., 2025)).

Qualitative comparisons. In Fig. 3, we present visual quality comparisons with several existing methods. SD3 fails to create a harmonious transition across the two contrastive regions, exhibiting a discontinuity in content (second row, an abrupt division between vibrant greenery and parched land). Flux.1-dev suffers from insufficient element and color contrast between the two regions, resulting in a visually indistinct composition (first row, both regions are dominated by similar grayish-brown hues). TextDiffuser-2 struggles to generate semantically consistent elements with text descriptions (second row, the obscured though well-substantial tree and blazing fire undermine the theme of "Save Forest"). OpenCOLE demonstrates suboptimal element arrangement and a compromised compositional structure (first row, children playing in a field incongruously are surrounded by a dull and blurred background). PosterGen struggles with visual text legibility since the colors of the visual text and the background are not easily distinguishable (second row, indistinct red visual texts are rendered on a fire background). PosterMaker generates chaotic and semantically incoherent content, damaging the overall aesthetic harmony (second row, the forest and fire elements are placed in a fragmented layout). In contrast, our results effectively balance element contrast and aesthetic harmony, while guaranteeing boundary coherence, producing visually striking and aesthetically appealing posters.

Quantitative comparisons. We present quantitative comparisons in Tab. 2, demonstrating that our method outperforms all compared methods across all metrics. ReContraster achieves highly aesthetically appealing results (LAS), exhibits visually striking element contrasts (RSD) with harmonious transitions across region boundaries (BGD), and accurately renders the visual texts (OCR). Additionally, ReContraster achieves top scores across all three VLM-based scores (CRE, VAI, and BIH).

Figure 4: Ablation study results with different variants of ReContraster.

User study. We conduct four user studies to evaluate whether ReContraster is preferred over state-of-the-art methods, using the same evaluation criteria introduced for the benchmark dataset. For each experiment, participants are shown a text description and seven generated posters. We conduct these experiments on Amazon Mechanical Turk (AMT), randomly selecting 100 samples from the benchmark dataset. Experiment results are polled by 25 volunteers independently. We present these scores in Tab. 2, highlighting the subjective advantages of our approach.

5.4 Ablation Study

We remove various proposed modules and designed losses to construct four baselines to evaluate their individual contributions. The evaluation scores and generated posters are shown in Tab. 2 and Fig. 4.

W/o Cognition & Arranger Interaction (CAI). We replace the cognition and arranger agents with a single LLM to eliminate interaction between these agents. As shown in Fig. 4 second row, this baseline generates posters with less compelling elements, resulting in a lower LAS score.

W/o Iterative Refiner Agent (IRA). We discard the refiner agent, thereby eliminating the iterative loop. As shown in Fig. 4 first row, this baseline reduces the contrast effect between elements, leading to a lower RSD score.

W/o Gradient Consistency Loss (GCL). We discard the gradient consistency loss during inference. As shown in Fig. 4 second row, this baseline results in visual discontinuities across the region boundary, leading to a worse BGD score.

W/o Joint Region Denoising (JRD). We remove the joint region denoising during inference. As shown in Fig. 4 first row, this baseline leads to disharmonious transitions across region boundaries, and reduces the LAS score.

Figure 5: Application scenarios of ReContraster.

5.5 Application

As shown in Fig. 5, ReContraster supports flexible integration of multi-lingual text descriptions, offers customized control over region divisions, and provides versatile editing of user-provided images[6].

6 Conclusion

In this paper, we introduce ReContraster, a training-free model that leverages the principle of "contrast effects" to produce visually striking and aesthetically appealing posters. ReContraster achieves element contrast, aesthetic harmony, and boundary coherence via the proposed compositional multi-agent system and hybrid regional denoising modules. We contribute a new benchmark dataset and showcase application scenarios. Quantitative results confirm its superiority over relevant methods.

[6]Technical details are provided in Sec. G.

Limitations

ReContraster performance is sensitive to user-specified region divisions, where overly small or highly complex divisions may diminish the visual appeal of the generated posters [7].

Acknowledgement

This work is supported by BUPT Innovation and Entrepreneurship Support Program (2025-YC-T029), National Natural Science Foundation of China (Grant No. 62136001), Beijing Major Science and Technology Project (Grant No. Z251100008125009), and the Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance. PKU-affiliated authors thank openbayes.com for providing computing resources.

References

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

Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. 2023. L-CAD: Language-based colorization with any-level descriptions using diffusion priors. In *Advances in Neural Information Processing Systems*.

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

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2024a. TextDiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and 1 others. 2024b. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. Seeclick: Harnessing gui grounding for advanced visual gui agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. 2006. Color harmonization. In *ACM SIGGRAPH Conference Papers*.

[7] Failure cases are presented in Sec. B

Yang Deng, Xuan Zhang, Wenxuan Zhang, Yifei Yuan, See-Kiong Ng, and Tat-Seng Chua. 2024. On the multi-turn instruction following for conversational web agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. 2022. Stytr2: Image style transfer with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and 1 others. 2021. Cogview: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems*.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*.

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

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hsiao Yuan Hsu, Xiangteng He, Yuxin Peng, Hao Kong, and Qing Zhang. 2023. PosterLayout: A new benchmark and approach for content-aware visual-textual presentation layout. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Naoto Inoue, Kento Masui, Wataru Shimoda, and Kota Yamaguchi. 2024. OpenCOLE: Towards reproducible automatic graphic design generation. *arXiv preprint arXiv:2406.08232*.

Peidong Jia, Chenxuan Li, Yuhui Yuan, Zeyu Liu, Yichao Shen, Bohan Chen, Xingru Chen, Yinglin Zheng, Dong Chen, Ji Li, and 1 others. 2023. COLE: A hierarchical generation framework for multi-layered and editable graphic design. *arXiv preprint arXiv:2311.16974*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *International Conference on Computer Vision*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

G Levkine. 2012. Prewitt, Sobel and Scharr gradient 5x5 convolution matrices. *Image Process. Articles, Second Draft*.

Zaijing Li, Gongwei Chen, Rui Shao, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing the emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.

Jiawei Lin, Jiaqi Guo, Shizhao Sun, Zijiang Yang, Jian-Guang Lou, and Dongmei Zhang. 2024. Layout-Prompter: Awaken the design ability of large language models. In *Advances in Neural Information Processing Systems*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Yunbo Long, Yuhan Liu, and Liming Xu. 2026. Emomas: Emotion-aware multi-agent system for high-stakes edge-deployable negotiation with bayesian orchestration. *arXiv preprint arXiv:2604.07003*.

Jian Ma, Yonglin Deng, Chen Chen, Haonan Lu, and Zhenyu Yang. 2025. Glyphdraw2: Automatic generation of complex glyph posters with diffusion models and large language models. In *Association for the Advancement of Artificial Intelligence*.

Anu MacIntosh-Murray. 2007. Poster presentations as a genre in knowledge communication: a case study of forms, norms, and values. *Science communication*.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Association for the Advancement of Artificial Intelligence*.

Zena O'Connor. 2015. Colour, contrast and gestalt theories of perception: The impact in contemporary visual communications design. *Color Research & Application*.

Jerry K Palmer and Jonathan S Gore. 2014. A theory of contrast effects in performance appraisal and social cognitive judgments. *Psychological Studies*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Eleanor Jiang, Chengfei Lv, and Huajun Chen. 2024. Autoact: Automatic agent learning from scratch for qa via self-planning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Laura D Scherer and Alan J Lambert. 2009. Contrast effects in priming paradigms: Implications for theory and research on implicit attitudes. *Journal of personality and social psychology*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*.

Jaejung Seol, Seojun Kim, and Jaejun Yoo. 2024. PosterLlama: Bridging design ability of langauge model to contents-aware layout generation. *arXiv preprint arXiv:2404.00995*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Rui Sun, Jie Ding, Chenghua Gong, Tianjun Gu, Yihang Jiang, Juyuan Zhang, Liming Pan, and Linyuan Lü. 2026. Topodim: One-shot topology generation of diverse interaction modes for multi-agent systems. *arXiv preprint arXiv:2601.10120*.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. 2024. AnyText: Multilingual visual text generation and editing. In *International Conference on Learning Representations*.

AW Utoyo, HD Aprilia, RADRI Kuntjoro-Jakti, and A Kurniawan. 2021. Visual communication design: Poster as an important way to encourage social distance in Jakarta when the epidemic 19. In *IOP conference series: earth and environmental science*.

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. 2023. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH Conference Papers*.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. 2024a. Genartist: Multimodal LLM as an agent for unified image generation and editing. In *Advances in Neural Information Processing Systems*.

Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. 2024b. StyleAdapter: A unified stylized image generation model. *International Journal of Computer Vision*.

Ziqing Wang, Yibo Wen, William Pattie, Xiao Luo, Weimin Wu, Jerry Yao-Chieh Hu, Abhishek Pandey, Han Liu, and Kaize Ding. 2025a. Polo: Preference-guided multi-turn reinforcement learning for lead optimization. *arXiv preprint arXiv:2509.21737*.

Ziqing Wang, Kexin Zhang, Zihan Zhao, Yibo Wen, Abhishek Pandey, Han Liu, and Kaize Ding. 2025b. A survey of large language models for text-guided molecular discovery: from molecule generation to optimization. *arXiv preprint arXiv:2505.16094*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Shuchen Weng, Peixuan Zhang, Zheng Chang, Xinlong Wang, Si Li, and Boxin Shi. 2023. Affective image filter: Reflecting emotions from text to images. In *ICCV*.

Meng Xu, Tong Zhang, Fuyun Wang, Yi Lei, Xin Liu, and Zhen Cui. 2024. MPDS: A movie posters dataset for image generation with diffusion model. *arXiv preprint arXiv:2410.16840*.

Tao Yang, Yingmin Luo, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. 2024. PosterLLaVa: Constructing a unified multi-modal layout generator with LLM. *arXiv preprint arXiv:2406.02884*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models.(2023). *arXiv preprint arXiv:2308.06721*.

Hui Zhang, Dexiang Hong, Tingwei Gao, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. 2025a. CreatiLayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. In *International Conference on Computer Vision*.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring collaboration mechanisms for llm agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. 2025b. Stage: Storyboard-anchored generation for cinematic multi-shot narrative. *arXiv preprint arXiv:2512.12372*.

Peixuan Zhang, Shuchen Weng, Jiajun Tang, Si Li, and Boxin Shi. 2025c. Towards deeper emotional reflection: Crafting affective image filters with generative priors. *IEEE transactions on pattern analysis and machine intelligence*.

Peixuan Zhang, Shuchen Weng, Chengxuan Zhu, Binghao Tang, Zijian Jia, Si Li, and Boxin Shi. 2026a. Affective image editing: Shaping emotional factors via text descriptions. *International Journal of Computer Vision*, 134(1):16.

Peixuan Zhang, Chang Zhou, Ziyuan Zhang, Hualuo Liu, Chunjie Zhang, Jingqi Liu, Xiaohui Zhou, Xi Chen, Shuchen Weng, Si Li, and Boxin Shi. 2026b. A benchmark and multi-agent system for instruction-driven cinematic video compilation. *arXiv preprint arXiv:2604.10456*.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2026c. Stable-rag: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. *arXiv preprint arXiv:2601.02993*.

Yunyao Zhang, Zikai Song, Hang Zhou, Wenfeng Ren, Yi-Ping Phoebe Chen, Junqing Yu, and Wei Yang. 2025d. $ga - s^3$: Comprehensive social network simulation with group agents. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Yunyao Zhang, Xinglang Zhang, Junxi Sheng, Wenbing Li, Junqing Yu, Yi-Ping Phoebe Chen, Wei Yang, and Zikai Song. 2025e. Semantic-aware logical reasoning via a semiotic framework. *arXiv preprint arXiv:2509.24765*.

Yiming Zhao and Zhouhui Lian. 2024. UDiffText: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. In *European Conference on Computer Vision*.

Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. 2019. Content-aware generative modeling of graphic design layouts. *ACM TOG*.

Min Zhou, Chenchen Xu, Ye Ma, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. 2022. Composition-aware graphic layout GAN for visual-textual presentation designs. In *International Joint Conference on Artificial Intelligence*.

Appendices

A Contrast Effects

As shown in Fig. 6, we present representative artworks from David Ambarzumjan's "Brushstrokes in Time" series. Ambarzumjan masterfully juxtaposes disparate elements to create a powerful visual impact: flourishing flowers with majestic waterscapes (Fig. 6 (a)), urban landscapes with vibrant nature (Fig. 6 (b)), and city streets with underwater worlds (Fig. 6 (c)). These examples illustrate how artists leverage the principle of "Contrast Effects" (Palmer and Gore, 2014; Scherer and Lambert, 2009; O'Connor, 2015) to create visually striking artworks. Inspired by this artistic approach, our model strategically utilizes user requests (including text descriptions of the theme and visual texts, and a mask indicating region divisions) to generate contrastive elements with enriched colors. It then organizes these elements into an aesthetically pleasing composition, ensuring harmonious transitions across regional boundaries to produce a visually striking and high-quality final poster.

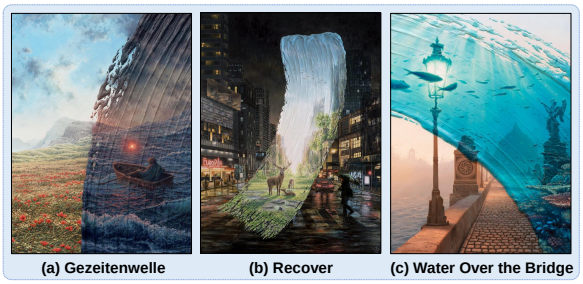

Figure 6: Selected artworks from David Ambarzumjan's "Brushstrokes in Time" series.

B Failure Cases

The performance of ReContraster in regional contrast poster generation relies on the user-provided region divisions. The model faces challenges when these divisions are either overly small or highly complex. As shown in Fig. 7 (left), an overly small region division (*i.e.*, a solitary dead tree against an expansive iceberg) results in limited contrast and reduced visual impact due to the lack of diverse elements within the small region. In addition, as shown in Fig. 7 (right), a highly complex region division (*i.e.*, an intricate winding path) causes the intended contrastive elements to become muddled, leading to an unclear central theme and reduced aesthetic appeal.

Figure 7: Visualization of failure cases.

C Robust Experiment

To evaluate the robustness of ReContraster's proposed compositional multi-agent system, we replace GPT-4o (Achiam et al., 2023) with Llama 3 (Dubey et al., 2024). As shown in Tab. 3 and Fig. 8, this baseline maintains comparable performance and demonstrates that the system is not dependent on a specific LLM.

Table 3: Robust experiment results.

Method	LAS ↑	RSD ↑	BGD ↓	OCR ↑
ReContraster-llama3	4.9049	803.10	**0.0352**	0.64
ReContraster-GPT-4o	**5.0966**	**842.60**	0.0375	**0.65**

Figure 8: Visualization of robust experiments.

D Model Efficiency

To evaluate the efficiency of our model, we measure the average time required to generate the poster on the test set. On average, ReContraster requires 2.8 attempts and a total of 295.9 seconds to generate a poster, where the diffusion process spends 117.6 seconds, with the time attributed to API scheduling and LLM reasoning. For multi-region scaling (3, 4, 5 regions), the inference times are approximately 338, 384, 417 seconds, respectively.

To fully address the trade-off between speed and quality, we conduct comprehensive ablations on the parameter to analyze the impact of various factors on efficiency and output quality. Specifically, we adjusted key parameters such as the number of denoising steps $t = 25, 50, 75$, the threshold step $\tau = 5, 10, 15$, and the solver configuration with the

Euler method or Heun's method. During comparison, we report results with $t = 50$, $\tau = 10$, and use the Euler method. The results are shown in Tab. 4.

Table 4: Ablation study on model efficiency and quality.

Method	LAS ↑	RSD ↑	BGD ↓	OCR ↑
$t = 25$	5.0370	705.53	0.0439	0.64
$t = 75$	5.0398	781.92	**0.0345**	0.65
$\tau = 5$	4.9956	808.18	0.0394	0.64
$\tau = 15$	5.0288	812.31	0.463	**0.66**
Heun solver	5.0951	821.01	0.0401	**0.66**
Ours	**5.0966**	**842.6**	0.0375	0.65

Method	CRE ↑	VAI ↑	BIH ↑	Time ↓
$t = 25$	7.79	7.64	6.80	**209.2**
$t = 75$	7.76	7.73	**7.17**	336.3
$\tau = 5$	**7.98**	7.49	6.96	267.7
$\tau = 15$	7.73	7.70	6.66	341.3
Heun solver	7.71	7.77	6.87	233.86
Ours	7.82	**7.87**	7.04	295.9

E Hyperparameter Setting

As described in Sec.3.3 of the main paper, our two-stage hybrid denoising strategy incorporates a ratio hyperparameter τ as a threshold for the steps in the gradient consistency loss and joint region denoising phases. We adjust this hyperparameter to explore its influence, visualizing the corresponding qualitative results in Fig. 9. We find that $\tau = 10$ effectively balances regional blending and boundary divisions. Consequently, we adopt this value as the default configuration for our method.

Figure 9: Qualitative results by adjusting hyperparameter τ in the hybrid denoising strategy.

F Evaluation Metrics Details

As illustrated in Sec.5.2 of the main paper, we employ the GPT-4o (Achiam et al., 2023) to assess the performance of relevant methods. For better transparency and reproducibility, we present the used prompts in Fig.10.

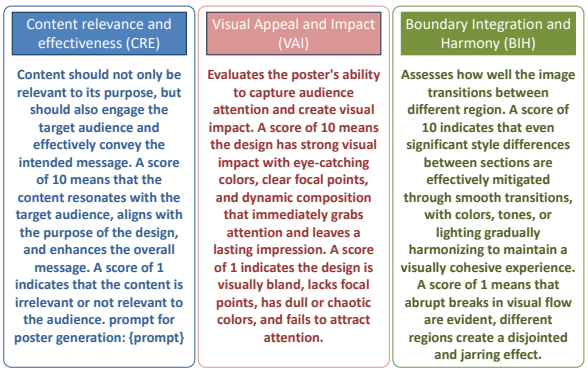

Figure 10: Evaluation prompts for CRE, VAI, and BIH.

G Application Technology Details

As illustrated in Sec.5.5 of the main paper, we present additional application scenarios of ReContraster, including commercial, event, decorative, travel, movie, and social advocacy posters. We further provide the corresponding technical details:

- **Flexible Integration:** We replace OpenCole (Inoue et al., 2024) with AnyText (Tuo et al., 2024) for text rendering, enabling support for Chinese characters. By freely integrating advanced visual text rendering models, ReContraster supports a wide range of languages (*e.g.*, French, German, Japanese, and Arabic).
- **Customized Control:** When a user provides a mask with multiple regions, the system adapts its process accordingly. First, the cognition agent is instructed to generate a unique contrastive element for each specified region. Subsequently, the arranger agent organizes these multiple elements into a layout based on the provided divisions. As a result, ReContraster is capable of handling posters with multi-region and irregular layouts.
- **Versatile Editing:** Given an existing image and text description for specified regions, ReContraster generates content for those regions while preserving the original region's content. Specifically, we first extract elements from the existing image and use them as prompts for generating corresponding contrastive elements. During the denoising process, the latent code of the original region is preserved, while only denoising the specified regions.

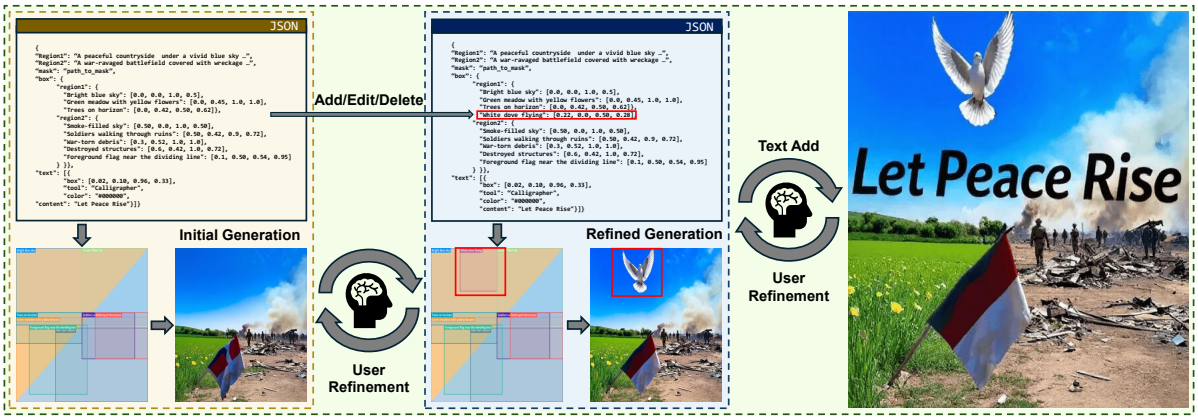

Figure 11: Illustration of real-time human involvement.

Figure 12: The results of font control.

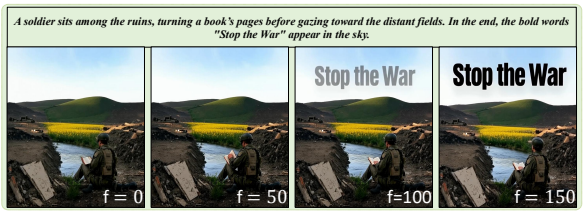

Figure 13: The results of modality scalability.

H Modular Flexibility and Interaction

Our proposed ReContraster is designed with a highly modular and interactive architecture, offering significant flexibility and controllability in the poster generation process. Specifically, this is demonstrated in the following three aspects:

- **Human Involvement:** To accommodate the need for precise customization, our system allows for real-time human intervention. Users can directly access and modify the intermediate JSON files generated by the arranger agent, as illustrated in Fig. 11.

- **Font Control:** Our modular design supports seamlessly integrating advanced text rendering modules (*e.g.*, AnyText and Calligrapher) to replace the default text rendering engine, as shown in Fig. 12.

- **Modality Scalability:** The capabilities of ReContraster extend beyond static images. By invoking existing image-to-video models after the image generation phase, users can easily animate the static designs into dynamic video posters, demonstrating strong modality scalability, as shown in Fig. 13.

I Additional Dataset Samples

As illustrated in Sec.4 of the main paper, we collect a benchmark dataset tailored for regional contrast poster generation. This dataset includes high-quality posters, corresponding text descriptions, and region division masks. Additional examples from this dataset are presented in Fig. 14 to inspire relevant research.

J Additional Application Samples

As illustrated in Sec.5.5 of the main paper, ReContraster is capable of creating diverse designs. To further showcase its wide-ranging applicability, we present additional poster examples in Fig. 15.

K Visualization of the Workflow

As shown in Fig. 16 and Fig. 17, we present a representative example where each stage of the ReContraster pipeline is executed correctly. This includes user input and information extraction, element identification and layout organization, image generation with hybrid denoising, and the final text rendering. This stage-by-stage process also shows a high level of interpretability.

Figure 14: Additional examples from our regional contrast poster dataset, showcasing diverse poster structures and thematic variations.

Figure 15: Additional application samples of ReContraster.

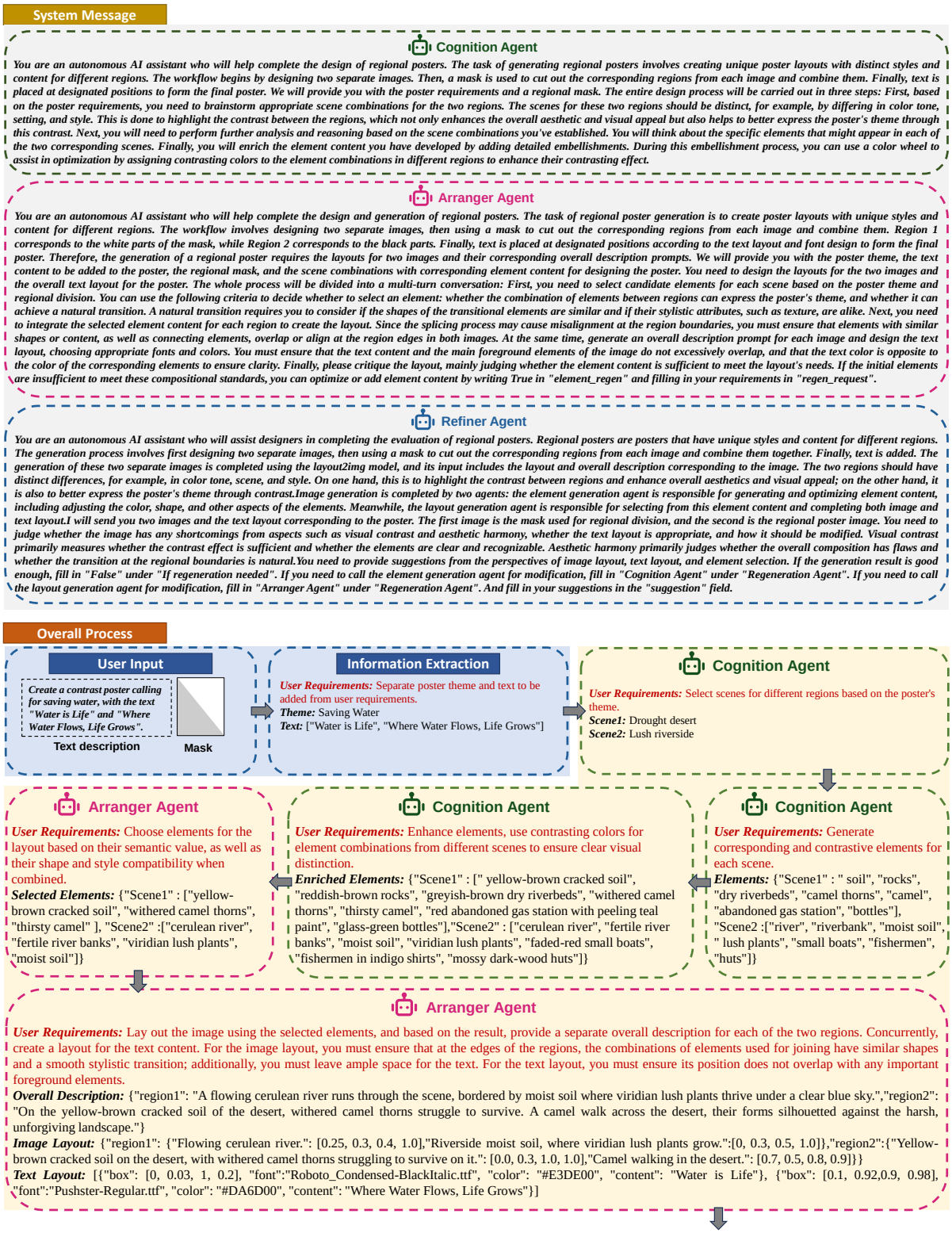

Figure 16: The workflow example of ReContraster (Part 1).

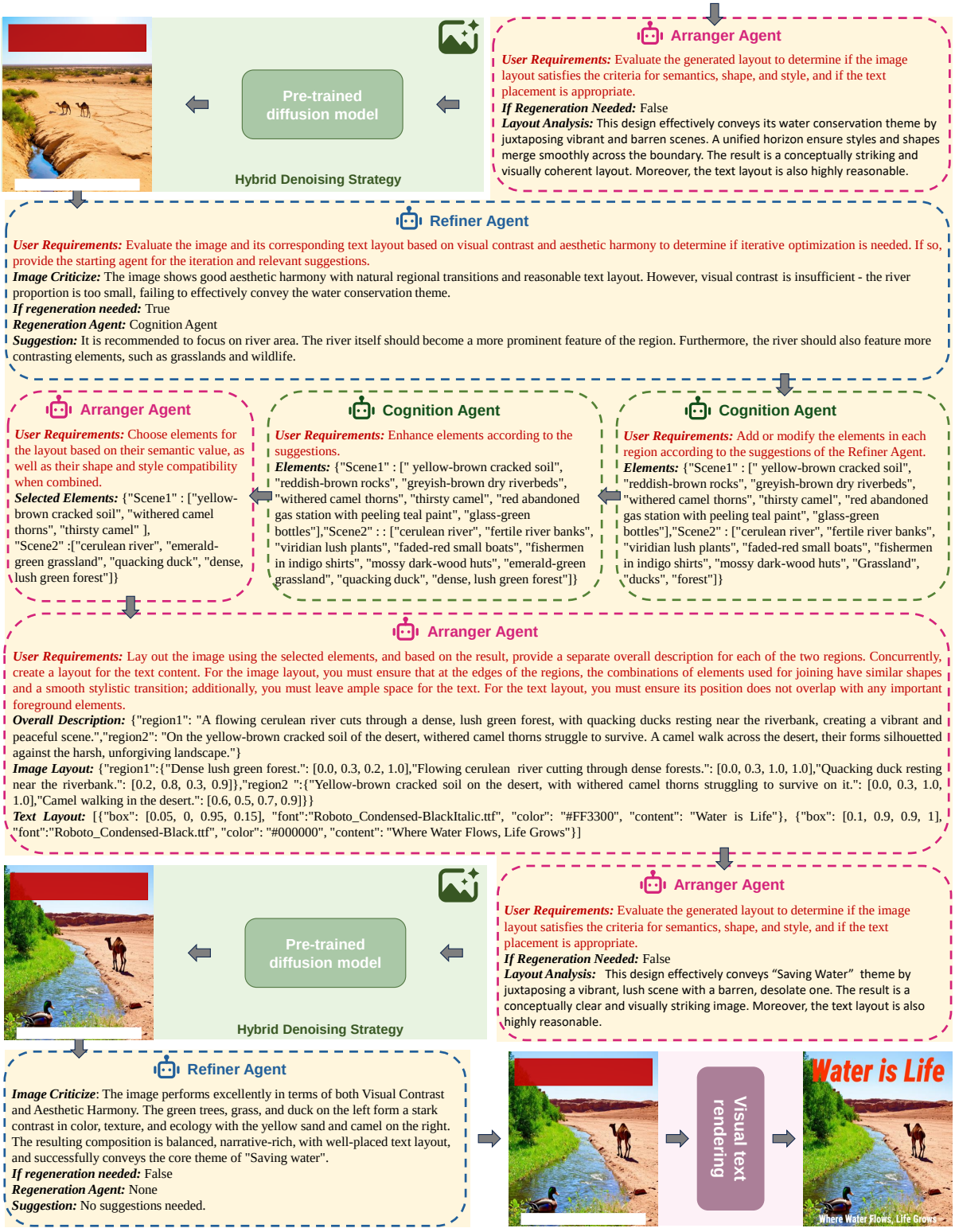

Figure 17: The workflow example of ReContraster (Part 2).