

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 DECOUPLING GLOBAL STRUCTURE AND LOCAL RE- FINEMENT: BLUEPRINT-GUIDED SCROLL GENERA- TION WITH DIRECT PREFERENCE OPTIMIZATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Existing methods for generating long scroll images, often fail to maintain global structural and stylistic consistency, resulting in artifacts like content repetition. To address this, we propose the Dual-Resolution Scroll Generation with Preference Optimization (DRSPO) framework. Our approach decouples global composition from local refinement by first generating a low-resolution (LR) blueprint to establish a coherent overall structure. This LR blueprint then guides a high-resolution (HR) feature to render fine-grained details. We further enhance generation quality by incorporating Direct Preference Optimization (DPO) at both stages, and we introduce a novel theoretical adaptation to apply preference tuning directly to the region-based generation process. Experimental results demonstrate that our method produces high-quality long scroll images with reasonable global structure and fine-grained details.

1 INTRODUCTION

The generation of long scroll images presents a challenging task within the field of image generation. Currently, mainstream image generation models primarily focus on producing images with an aspect ratio of 1:1. Directly applying these models like SDXL (Podell et al., 2023) or Flux (Labs, 2024) to generate high aspect ratio images encounters the Out-of-Distribution (OOD) problem, leading to a significant degradation in generation quality.

A prevalent strategy for high-resolution long scroll generation involves region-based methods, epitomized by MultiDiffusion (Bar-Tal et al., 2023). This approach is particularly effective at rendering high-fidelity local details by processing the image in smaller, manageable regions. Building on this foundation, subsequent works have introduced targeted improvements: Merge-Attend-Diffuse (MAD) operator (Quattrini et al., 2024) enhances semantic coherence by merging features across different window views, SyncDiffusion (Lee et al., 2023) focuses on improving stylistic consistency between these windows, and ElasticDiffusion (Haji-Ali et al., 2024) further refines overall local coherence. Conversely, an alternative paradigm involves the direct end-to-end application of powerful base models like SDXL (Podell et al., 2023) and Flux (Labs, 2024). This approach demonstrates a distinct advantage in establishing global coherence and stylistic consistency across the entire scroll.

Despite these advancements, both paradigms suffer from some limitations. The core deficiency of region-based methods is the absence of a holistic global plan. This frequently manifests as severe artifacts, most notably the unnatural repetition of objects, as illustrated in Fig. 1. While the direct application of base models provides more robust global control, their ability to render fine-grained details does not match the fidelity achieved by region-based techniques. Consequently, existing works present a fundamental tradeoff between global structural coherence and fine-grained local refinement.

To address these challenges, we propose the Dual-Resolution Scroll Generation with Preference Optimization (DRSPO) framework. Our framework is comprised of two core components. The first is a Dual-Resolution Generation Pipeline (detailed in Sec.3.1), which utilizes a low-resolution (LR) model to generate a blueprint that establishes the global compositional structure. This LR blueprint subsequently provides robust guidance for the high-resolution (HR) generation stage. The second component is a novel Direct Preference Optimization (DPO) method tailored for the MultiDiffusion



095 Figure 1: Scrolls(1024×4096) generated using prompts: “Paint a scene of a canal-laced city, like
096 Venice or Amsterdam. Gondolas or boats glide through the waterways, passing under arched bridges
097 and alongside colorful, historic houses that seem to rise directly from the water.” As can be observed,
098 the images generated by our method exhibit more reasonable and coherent layouts. Flux+ESRGAN
099 also generates well-defined results. In contrast, other methods, which lack guidance from global
100 information, produce images with severe object repetition.

101
102 process(detailed in Sec.3.2). It directly fine-tunes the base generator to enhance the fine-grained
103 details within each local region. The efficacy of our framework is demonstrated by state-of-the-
104 art performance on metrics such as HPS v2 and ImageReward. Visually, Fig. 1 confirms that our
105 method excels at synthesizing coherent global structures with high-fidelity local details.

106 Our work introduces a Dual-Resolution Scroll Generation Method that enables controlled generation
107 of the global structure, ensuring the final long scroll possesses both a coherent global composition

108 and consistent style. Furthermore, we adapt the DPO method for application within the MultiDiffusion
 109 scenario, specifically to optimize the generation of long scroll images. The effectiveness of this
 110 combined approach is demonstrated by our experimental results, which show our method’s ability to
 111 generate high-quality outputs and achieve leading scores on key metrics for human preference and
 112 semantic fidelity.

114 2 PRELIMINARY

116 2.1 MULTIDIFFUSION

118 MultiDiffusion (Bar-Tal et al., 2023) is a training-free method that adapts a pre-trained diffusion
 119 model for long scroll generation using a sliding window approach. It is built upon a base model Φ
 120 whose standard denoising step is formulated as:

$$121 \quad \Phi : \mathcal{I} \times \mathcal{Y} \rightarrow \mathcal{I}, \mathcal{I} \in \mathbb{R}^{H \times W \times C} \quad (1)$$

123 where Φ maps a noisy latent from the image space \mathcal{I} to a denoised one under conditions from \mathcal{Y} ,
 124 such that $I_{t-1} = \Phi(I_t|y)$. To generate a long scroll image $\mathcal{J} \in \mathbb{R}^{H' \times W' \times C}$, MultiDiffusion defines
 125 a new process Ψ :

$$126 \quad \Psi : \mathcal{J} \times \mathcal{Z} \rightarrow \mathcal{J} \quad (2)$$

127 The new diffusion step, $J_{t-1} = \Psi(J_t|y)$, is constrained by the base model Φ . This link is established
 128 via a mapping F_i that extracts the i -th window and its inverse F_i^{-1} that maps it back. The objective
 129 is to align this new process with the pre-trained model by solving the following optimization at each
 130 step:

$$131 \quad \Psi(J_t|y) = \arg \min_{J \in \mathcal{J}} \sum_{i=1}^n \|F_i(J) - \Phi(F_i(J_t), y_i)\| \quad (3)$$

134 In practice, this optimization is implemented by denoising each window individually and then com-
 135 posing the complete long scroll J_{t-1} via a weighted average of the results:

$$136 \quad J_{t-1} = \frac{\sum_i F_i^{-1}(W_i \Phi(F_i(J_t), y_i))}{\sum_i F_i^{-1}(W_i)} \quad (4)$$

139 2.1.1 DPO-DIFFUSION

141 Direct Preference Optimization (DPO) (Rafailov et al., 2023; Wallace et al., 2024) is a method that
 142 fine-tunes a model to align with preferences by directly optimizing a policy against a reward function
 143 implicitly defined by a preference dataset. In the context of diffusion models, DPO-Diffusion defines
 144 this reward over the entire denoising trajectory:

$$145 \quad r(c, x_0) = \mathbb{E}_{p_\theta(x_{1:T}|x_0, c)}[R(c, x_{0:T})] \quad (5)$$

146 The reinforcement learning (RL) objective, using the KL divergence as a constraint, is formulated
 147 as:

$$148 \quad \max_{p_\theta} \mathbb{E}_{c \sim \mathcal{D}_c, x_{0:T} \sim p_\theta(x_{0:T}|c)}[r(c, x_0)] - \beta \mathbb{D}_{KL}[p_\theta(x_{0:T}|c) \parallel p_{\text{ref}}(x_{0:T}|c)] \quad (6)$$

150 By substituting the difference in rewards for preferred (x_0^w) and dispreferred (x_0^l) samples from a
 151 preference dataset \mathcal{D} , the final DPO-Diffusion loss function is derived as:

$$153 \quad L_{\text{DPO-Diffusion}} = -\mathbb{E}_{(c, x_0^w, x_0^l) \sim \mathcal{D}} \log \sigma \left(\beta \mathbb{E}_{x_{1:T}^w \sim p_\theta(\cdot|x_0^w, c), x_{1:T}^l \sim p_\theta(\cdot|x_0^l, c)} \right. \\ \left. \left[\log \frac{p_\theta(x_{0:T}^w|c)}{p_{\text{ref}}(x_{0:T}^w|c)} - \log \frac{p_\theta(x_{0:T}^l|c)}{p_{\text{ref}}(x_{0:T}^l|c)} \right] \right) \quad (7)$$

157 3 METHOD

160 Existing high-resolution long scroll generation methods face challenges with global structure and
 161 content repetition. Approaches like MultiDiffusion (Bar-Tal et al., 2023), which rely on “divide
 and conquer” strategy of independently denoising and then aggregating image regions, inherently

lack the global context necessary for structural coherence. While the end-to-end methods lack local refinement. These limitations highlight the need for an external control mechanism to impose global structural consistency while maintaining the local details.

To address these issues, we propose the Dual-Resolution Scroll Generation with Preference Optimization (DRSPO) framework. Our method first uses a Low-Resolution (LR) generator to establish a coherent global blueprint, which then guides a High-Resolution (HR) generator to render fine-grained details. We further enhance this pipeline by integrating DPO (Wallace et al., 2024) at both stages, fine-tuning the models on preference data to improve generation quality, resulting in scrolls that are both globally coherent and locally detailed.

3.1 GENERATION PIPELINE

Our framework is implemented as a two-stage generation pipeline: the Low-Resolution blueprint generation stage and the subsequent High-Resolution detail rendering stage. The framework of our generation pipeline is shown in Fig. 2.

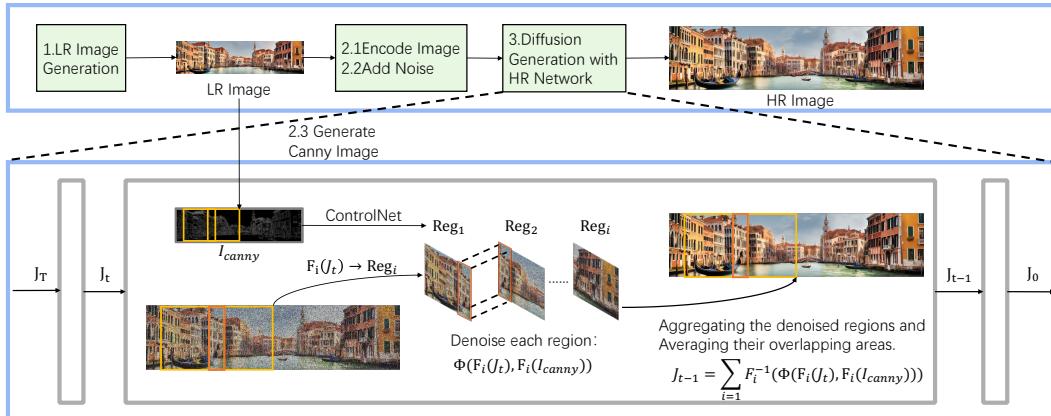


Figure 2: Detailed schematic of our Dual-Resolution Generation Pipeline. The process begins with the generation of a Low-Resolution (LR) image (Step 1), which serves as a global ‘blueprint’ for HR generation (Step 2.1-2.3). The lower panel provides an in-depth view of a single denoising step within the HR generator (Step 3).

3.1.1 LOW-RESOLUTION IMAGE GENERATION

The primary objective of this stage is to establish the global structure and style for the final long scroll. To circumvent the content repetition and structural control issues inherent in direct high-resolution synthesis, we first employ a pre-trained base model to generate the complete long scroll at a lower resolution. This process can be formalized as:

$$\Phi_{\text{LR}} : \mathcal{I} \times \mathcal{Y} \rightarrow \mathcal{I}, \mathcal{I} \in \mathbb{R}^{H_1 \times W_1 \times C}, I_{\text{LR}} = \Phi_{\text{LR}}(z, c) \quad (8)$$

where $I_{\text{LR}} \in \mathbb{R}^{H_1 \times W_1 \times C}$ is the generated LR image, and Φ_{LR} is the pre-trained model.

To further enhance the quality and consistency of the LR image, we fine-tune the model Φ_{LR} using DPO (Wallace et al., 2024). We construct a preference dataset by evaluating a corpus of long scroll images with an Aesthetic Score (Schuhmann et al., 2022), from which we form preference pairs. The detailed methodology for dataset construction is described in Sec. 5.1.1. The DPO fine-tuning process is guided by the following loss function:

$$L(\theta) = -\mathbb{E}_{t, J_t^w \sim q(J_t | J_0^w), J_t^l \sim q(J_t | J_0^l)} \log \sigma \left((-\beta T \omega(\lambda_t)) \left((\| \epsilon_t^w - \epsilon_\theta^w \| - \| \epsilon_t^w - \epsilon_{\text{ref}}^w \|) - (\| \epsilon_t^l - \epsilon_\theta^l \| - \| \epsilon_t^l - \epsilon_{\text{ref}}^l \|) \right) \right) \quad (9)$$

216 3.1.2 HIGH-RESOLUTION SCROLL GENERATION
217

218 The High-Resolution generation stage employs a modified MultiDiffusion framework, where the
219 Low-Resolution (I_{LR}) image from the preceding stage serves as a multi-source guidance signal.
220 This approach ensures the final long scroll is rich in detail while strictly adhering to the established
221 global structure.

222 Rather than using standard Gaussian noise, we initialize the process with the latent features of the
223 LR image, obtained via an image encoder \mathbf{E} :

$$224 z_{\text{init}} = \mathbf{E}(I_{LR}), J = \Psi(z_{\text{init}}, \mathbf{c}) \quad (10)$$

225 In this formulation, $J \in \mathbb{R}^{H_2 \times W_2 \times c}$ is the final generated high-resolution long scroll. To ensure
226 structural alignment, we maintain an identical aspect ratio k for both HR and LR images, such that
227 $\frac{W_1}{H_1} = \frac{W_2}{H_2} = k$.
228

229 **ControlNet** To impose structural control, the LR image is first converted into a line art represen-
230 tation, which guides the HR generation. Within the MultiDiffusion process, where each window is
231 generated independently, ControlNet (Zhang et al., 2023) provides localized control for each corre-
232 sponding region, as denoted by the mapping function $F_i, i \in \{1, \dots, n\}$.

$$233 I_{\text{canny}} = \text{Canny}(I_{LR}) \\ 234 I_{t-1}^i = \Phi(F_i(J_t), F_i(I_{\text{canny}}), \mathbf{c}) \quad (11)$$

236 3.2 APPLYING DPO TO MULTIDIFFUSION
237

238 In MultiDiffusion (Bar-Tal et al., 2023), each denoising step involves local denoising followed by
239 inter-region fusion. Consequently, deriving the noisy latent J_t at an arbitrary timestep t from the
240 scroll J_0 is non-trivial. To align with the DPO-Diffusion (Wallace et al., 2024) derivation, we ap-
241 proximate the forward process $q(J_t|J_0)$ by applying noise directly to the entire image (Ho et al.,
242 2020; Song et al., 2020):

$$243 q(J_t|J_0) = \mathcal{N}(J_t; \sqrt{\alpha_t}J_0, (1 - \alpha_t)\mathbf{I}) \quad (12)$$

244 We then define the reward function on the complete long scroll image:

$$245 r(c, J_0) = \mathbb{E}_{p_\theta(J_{1:T}|J_0, c)}[R(c, J_{0:T})] \quad (13)$$

246 Maximizing this reward while constraining the policy via KL divergence gives:

$$247 \max_{p_\theta} \mathbb{E}_{c \sim \mathcal{D}_c, J_{0:T} \sim p_\theta(J_{0:T}|c)}[r(c, J_0)] - \beta \mathbb{D}_{KL}[p_\theta(J_{0:T}|c) \parallel p_{\text{ref}}(J_{0:T}|c)] \quad (14)$$

249 Solving for the reward function yields:

$$250 R(c, J_{0:T}) = \beta \log \frac{p_\theta^*(J_{0:T}|c)}{p_{\text{ref}}(J_{0:T}|c)} + \beta \log Z(c) \\ 251 r(c, J_0) = \beta \mathbb{E}_{p_\theta(J_{1:T}|J_0, c)} \left[\log \frac{p_\theta^*(J_{0:T}|c)}{p_{\text{ref}}(J_{0:T}|c)} \right] + \beta \log Z(c) \quad (15)$$

254 The DPO loss function for long scroll generation is thus:

$$255 L_{\text{DPO-MultiDiffusion}} = -\mathbb{E}_{(c, J_0^w, J_0^l) \sim \mathcal{D}} \log \sigma \\ 256 \left(\beta \mathbb{E}_{J_{1:T}^w \sim p_\theta(J_{1:T}|J_0^w, c), J_{1:T}^l \sim p_\theta(J_{1:T}|J_0^l, c)} \left[\log \frac{p_\theta(J_{0:T}^w|c)}{p_{\text{ref}}(J_{0:T}^w|c)} - \log \frac{p_\theta(J_{0:T}^l|c)}{p_{\text{ref}}(J_{0:T}^l|c)} \right] \right) \quad (16)$$

259 Since the reverse trajectory $p_\theta(x_{1:T}|x_0)$ is intractable, we follow the approach in DPO and substitute
260 it with the forward process $q(x_{1:T}|x_0)$, using the simplified noising process defined in Equation. 12.
261 By applying Jensen's inequality, we can derive the following upper bound:

$$262 L_{\text{DPO-MultiDiffusion}} \leq -\mathbb{E}_{t, J_t^w \sim q(J_t|J_0^w), J_t^l \sim q(J_t|J_0^l)} \log \sigma \left(-\beta T \left(\right. \right. \\ 263 \left. \left. + \mathbb{D}_{KL}(q(J_{t-1}^w|J_0^w, J_t^w) \parallel p_\theta(J_{t-1}^w|J_t^w, c)) \right. \right. \\ 264 \left. \left. - \mathbb{D}_{KL}(q(J_{t-1}^w|J_0^w, J_t^w) \parallel p_{\text{ref}}(J_{t-1}^w|J_t^w, c)) \right. \right. \\ 265 \left. \left. - \mathbb{D}_{KL}(q(J_{t-1}^l|J_0^l, J_t^l) \parallel p_\theta(J_{t-1}^l|J_t^l, c)) \right. \right. \\ 266 \left. \left. + \mathbb{D}_{KL}(q(J_{t-1}^l|J_0^l, J_t^l) \parallel p_{\text{ref}}(J_{t-1}^l|J_t^l, c)) \right) \right) \quad (17)$$

270 In the MultiDiffusion framework, the mean of the denoised scroll is an aggregation of the means
 271 from each local window:

$$273 \quad p(J_{t-1}|J_t, c) \approx \mathcal{N} \left(J_{t-1}; \frac{1}{n} \sum_{i=1}^n F_i^{-1}(\mu_{\theta,i}(J_t, t, c)), \Sigma_t \right) \quad (18)$$

275 where $\mu_{\theta,i}(J_t, t, c)$ is the mean predicted by the local denoiser Φ for window i . The loss can be
 276 expressed in terms of the predicted noise ϵ yields:
 277

$$278 \quad L(\theta) = -\mathbb{E}_{t, J_t^w \sim q(J_t|J_0^w), J_t^l \sim q(J_t|J_0^l)} \log \sigma \left((-\beta T \omega(\lambda_t)) (\right. \\ 279 \quad \left. \| \epsilon^w - \sum_{i=1}^n F_i^{-1}(\epsilon_{\theta}(x_t^w, t)) \| - \| \epsilon^w - \sum_{i=1}^n F_i^{-1}(\epsilon_{\text{ref}}(x_t^w, t)) \| \right. \\ 280 \quad \left. - (\| \epsilon^l - \sum_{i=1}^n F_i^{-1}(\epsilon_{\theta}(x_t^l, t)) \| - \| \epsilon^l - \sum_{i=1}^n F_i^{-1}(\epsilon_{\text{ref}}(x_t^l, t)) \|) \right) \quad (19)$$

286 where $x_t^w = F_i(J_t^w)$ and $x_t^l = F_i(J_t^l)$ are the noisy latents for the respective windows.
 287

288 3.3 DISCUSSION

290 3.3.1 VARIANCE CONSIDERATIONS

291 In the derivation of Equation. 18, our analysis simplified the process by disregarding the effect of
 292 local window aggregation on the variance of the complete long scroll. However, as discussed in (Sun
 293 et al., 2025), the window stitching in MultiDiffusion has a non-trivial impact on variance. Assuming
 294 N windows overlap in a given region, the mean and variance of the denoised latent x_{t-1} in that
 295 region are:

$$296 \quad x_{t-1} \sim \mathcal{N} \left(\frac{\sum_i^N \mu_{t,i}}{N}, \frac{\sigma_t^2}{N} \right) \quad (20)$$

299 We now re-examine the derivation from Equation. 17 under this condition. The original derivation
 300 was based on the KL divergence between two Gaussian distributions with identical covariance $\Sigma =$
 301 $\sigma^2 \mathbf{I}$:

$$302 \quad D_{\text{KL}}(\mathcal{N}(\mu_1, \Sigma) || \mathcal{N}(\mu_2, \Sigma)) \\ 303 \quad = \frac{1}{2} \left(\text{tr}(\Sigma^{-1} \Sigma) + (\mu_2 - \mu_1)^T \Sigma^{-1} (\mu_2 - \mu_1) - d + \log \frac{\det \Sigma}{\det \Sigma} \right) = \frac{1}{2\sigma^2} \|\mu_1 - \mu_2\|_2^2 \quad (21)$$

306 In the MultiDiffusion context, however, the effective variance is scaled by $1/N$. The KL divergence
 307 between two Gaussians (Hershey & Olsen, 2007) with differing variances (σ_1^2 and σ_2^2) is therefore
 308 more appropriate:

$$310 \quad D_{\text{KL}}(P_1 || P_2) = \frac{1}{2} \left[2d \log \frac{\sigma_2}{\sigma_1} - d + d \frac{\sigma_1^2}{\sigma_2^2} + \frac{1}{\sigma_2^2} \|\mu_1 - \mu_2\|_2^2 \right] \\ 311 \quad = \frac{1}{2\sigma_2^2} \|\mu_1 - \mu_2\|_2^2 + \frac{1}{2} \left(2d \log \frac{\sigma_2}{\sigma_1} - d + \frac{d\sigma_1^2}{\sigma_2^2} \right) \quad (22)$$

314 Crucially, since the ratio of the variances $\frac{\sigma_2}{\sigma_1}$ is a constant in the MultiDiffusion scenario, this more
 315 accurate KL divergence formulation differs from the original only by an additive constant term. A
 316 term that is independent of the model parameters θ does not affect the location of the optima. There-
 317 fore, this variance consideration does not alter the DPO optimization process, and our previously
 318 derived loss function remains valid.
 319

320 3.3.2 LIMITATIONS

322 To apply the DPO framework to the MultiDiffusion process, we made a simplifying assumption
 323 about the forward noising process (Equation. 12), which may not perfectly align with the true dy-
 324 namics of MultiDiffusion’s region-based aggregation.

324 Furthermore, our data selection strategy, which constructs preference pairs by comparing the holistic
 325 quality of entire long scrolls, could be refined. The overall quality of a long scroll does not always
 326 correlate with the quality of its constituent regions; for example, a globally dispreferred image may
 327 still contain locally high-quality regions. A promising direction for future work would be to adopt a
 328 hybrid training strategy that simultaneously applies DPO at both the global long scroll level and the
 329 separate region level. This could be implemented via a composite loss function:

$$330 \quad L = L_1(\theta) + \lambda L_2(\theta) \quad (23)$$

331 where $L_1(\theta)$ is the long scroll DPO loss from Equation. 19, and $L_2(\theta)$ is a DPO loss function trained
 332 on individual image regions. The λ would balance two objectives.
 333

334 4 RELATED WORK

336 4.1 HIGH-ASPECT-RATIO IMAGE GENERATION

338 Standard text-to-image diffusion models, primarily trained on square images, struggle to generate
 339 high-aspect-ratio content, often producing repetitive or incoherent structures when tasked with out-
 340 of-distribution sizes. To overcome this, region-based, sliding-window approaches have become the
 341 dominant paradigm. MultiDiffusion (Bar-Tal et al., 2023) pioneered this area by denoising overlap-
 342 ping regions on a large canvas and blending them to ensure local coherence.

343 Building on this foundation, subsequent research has aimed to mitigate the artifacts inherent in this
 344 local-to-global process. For instance, Merge-Attend-Diffuse (MAD) operator (Quattrini et al., 2024)
 345 enhances long-range semantic consistency by sharing features across non-adjacent windows, while
 346 SyncDiffusion (Lee et al., 2023) and ElasticDiffusion (Haji-Ali et al., 2024) focus on improving
 347 stylistic uniformity and transitional smoothness. More recent methods like TwinDiffusion (Zhou
 348 & Tang, 2024) and SpotDiffusion (Frolov et al., 2024) have sought to further refine quality and
 349 optimize the efficiency of this region-based pipeline. Native high-resolution generators, such as
 350 the Transformer-based PixArt- α (Chen et al., 2023) and PixArt- Σ (Chen et al., 2024), leverage effi-
 351 cient DiT architectures and high-quality 4K training data to directly synthesize high-fidelity images.
 352 Alternatively, cascaded pipelines employ a "generate-then-upscale" strategy, utilizing foundation
 353 models (e.g., Stable Diffusion) followed by super-resolution networks like Real-ESRGAN (Wang
 354 et al., 2021) to enhance resolution. However, despite these incremental improvements, all such
 355 methods lack a dedicated mechanism for high-level compositional planning. This fundamental gap
 356 often leads to images that are locally seamless but globally repetitive or structurally flawed. Our
 357 work directly addresses this by introducing a global planning stage via a dual-resolution framework.

358 4.2 PREFERENCE OPTIMIZATION IN GENERATIVE MODELS

360 Aligning generative models with human preferences, such as aesthetic quality and prompt fidelity, is
 361 a critical challenge. Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022)
 362 was a foundational approach, where a separate reward model is trained on human preference data
 363 to guide the generator. However, RLHF is often complex and unstable to train. A more direct and
 364 stable alternative, DPO (Rafailov et al., 2023), has recently gained prominence. DPO bypasses the
 365 need for an explicit reward model by reframing the objective as a simple classification loss on pre-
 366 ferred and dispreferred samples, fine-tuning the policy directly. This technique has been successfully
 367 adapted for diffusion models in works like Diffusion-DPO (Wallace et al., 2024), proving effective
 368 at enhancing image quality and alignment. In our framework, we apply DPO not only to enhance
 369 global structural control during the planning stage but also extend the method itself to optimize the
 370 region-based MultiDiffusion process.

371 5 EXPRIMENTS

372 5.1 SETTINGS

373 5.1.1 PREFERENCE DATASET CONSTRUCTION

375 To facilitate Direct Preference Optimization (DPO) (Wallace et al., 2024), we constructed a large-
 376 scale preference dataset for long scroll images. The foundation of this dataset is a curated set of

378 4,936 high-quality prompts, which were derived from the corpus of (Zhang et al., 2024a). We use
 379 Coze to build a LLM-based workflow that filtered for outdoor scenes and textually augmented the
 380 content. Following the methodology of (Zhang et al., 2024b) to ensure sample diversity, we gener-
 381 ated a comparison group of six distinct images for each prompt using three models (MultiDiffusion
 382 (Bar-Tal et al., 2023), Stable Diffusion XL (Podell et al., 2023), and MAD (Quattrini et al., 2024))
 383 with two random seeds. Subsequently, every generated image was evaluated using aesthetic score
 384 (Schuhmann et al., 2022). By ranking the images within each comparison group based on these
 385 scores, we systematically established preference hierarchies, allowing us to extract the “preferred”
 386 and “dispreferred” sample pairs essential for aligning our models with desired aesthetic and struc-
 387 tural qualities through DPO training.

388 5.1.2 HYPERPARAMETERS 389

390 For DPO training, we use Adafactor to save memory. We train our model on single NVIDIA H800
 391 GPU using batch size of 1 pair. We train at resolution 1024×4096 for both stages. And for HR
 392 model, we use stride=32 and window size=128, corresponding to the resolution 1024×4096 . For
 393 both stage, we use Low rank adaptation(LoRA) to save memory and the rank is set to 64. We use
 394 learning rate=1e-6 with 1000 warm up steps. Following DPO, we use $\beta = 5000$ during training.

395 5.1.3 EVALUATION 396

397 For our evaluation, we constructed a test set from two sources: 220 new prompts for outdoor scenes
 398 generated by Google Gemini, and a hold-out set of 200 prompts from our training distribution that
 399 were not used during training. We then employed several long scroll generation methods to produce
 400 images for each prompt, benchmarking our approach against prominent baseline methods including
 401 MultiDiffusion (Bar-Tal et al., 2023), SDXL (Podell et al., 2023), MAD (Quattrini et al., 2024),
 402 ElasticDiffusion (Haji-Ali et al., 2024), SyncDiffusion (Lee et al., 2023), and the FLUX.1-dev (Labs,
 403 2024) version. The resulting images were assessed using a suite of quantitative metrics. To evaluate
 404 overall image quality and aesthetic appeal, we utilized three established scoring models: Aesthetic
 405 Score (Schuhmann et al., 2022), HPS v2 (Wu et al., 2023), ImageReward (Kirstain et al., 2023) and
 406 PickScore (Kirstain et al., 2023). We used the CLIP score (Hessel et al., 2021) to measure the
 407 semantic consistency between the generated images and their corresponding prompts.

408 5.2 QUANTITATIVE RESULTS 409

410 Table 1: Quantitative results of scrolls generated by different models. The numbers in parentheses
 411 indicate the specific rank within the group (lower is better). The 1st, 2nd and 3rd best values
 412 are highlighted with background colors. The **Average Rank** column reflects the overall balanced
 413 performance.

415 Model	Aesthetic Score \uparrow	HPS v2 \uparrow	PickScore \uparrow	ImageReward \uparrow	CLIP Score \uparrow	Average Rank \downarrow
416 MultiDiffusion	5.616 (8)	0.226 (4)	0.070 (7)	0.224 (5)	0.317 (2)	5.2
417 SDXL	5.921 (4)	0.190 (8)	0.076 (6)	-0.318 (8)	0.310 (4)	6.0
418 FLUX	4.994 (10)	0.175 (10)	0.027 (9)	-1.084 (9)	0.270 (10)	9.6
419 MAD	6.032 (1)	0.195 (7)	0.105 (1)	0.100 (6)	0.317 (2)	3.4
420 SyncDiffusion	5.744 (6)	0.197 (6)	0.068 (8)	-0.087 (7)	0.310 (4)	6.2
421 ElasticDiffusion	5.422 (9)	0.186 (9)	0.025 (10)	-1.351 (10)	0.273 (9)	9.4
422 FLUX + ESRGAN	5.922 (3)	0.235 (2)	0.097 (2)	0.236 (4)	0.306 (8)	3.8
423 SD 3.5 + ESRGAN	5.759 (5)	0.221 (5)	0.092 (4)	0.258 (3)	0.309 (6)	4.6
PixArt	6.016 (2)	0.236 (1)	0.092 (5)	0.397 (1)	0.308 (7)	3.2
424 DRSPR (Ours)	5.726 (7)	0.228 (3)	0.096 (3)	0.275 (2)	0.321 (1)	3.2

425 The quantitative results clearly demonstrate the superiority of our proposed method. Our model
 426 significantly outperforms competing approaches on the metrics most aligned with image quality
 427 preference and semantic fidelity, achieving the highest scores in HPS v2, ImageReward, and CLIP
 428 Score, along with second-place in PickScore. This robust performance validates the effectiveness of
 429 our DPO-based pipeline in producing visually appealing, compositionally sound, and semantically
 430 coherent images. Notably, while our method does not achieve the top rank on the Aesthetic Score,
 431 we attribute this to the metric’s well-documented bias towards favoring high-frequency local textures
 and details over global structural coherence.

432 5.3 QUALITATIVE RESULTS
433462 Figure 3: Further qualitative comparisons between our method and other baselines.
463

464 We provide more comparisons with other baselines as shown in Fig. 3. Baseline methods frequently
465 suffer from object repetition. For instance, in the first row, other approaches generate multiple roads,
466 disrupting the scene’s logical coherence. Our method successfully eliminates the repetitive artifacts
467 that plague competing approaches, achieving better global structural coherence and fine-grained
468 details.
469

470 5.4 ABLATION STUDY
471

472 We conduct an ablation study to validate the contributions of our key components: Low-Resolution
473 blueprint guidance and DPO training for high-resolution generation. The results in Table. 2 reveal
474 their distinct roles. The removal of either component leads to a notable degradation in the final out-
475 put. Specifically, omitting DPO training primarily harms human preference scores, whereas omitting
476 the LR guidance causes a sharp HPS v2 score decline. This confirms their complementary roles.
477 While the results of removing control blocks and DPO blocks both score higher on the Aesthetic
478 Score, we attribute this to the metric’s known bias towards local textures, which fails to penalize its
479 global incoherence. Furthermore, integrating an additional control mechanism (IP-Adapter) proves
480 detrimental, degrading performance across all metrics.
481

482 6 CONCLUSION
483

484 In this work, we propose the Dual-Resolution Scroll Generation with Preference Optimization (DR-
485 SPO) framework. Our method decouples global composition from local refinement by using a low-
486 resolution blueprint to guide a region-based high-resolution model. A key contribution is our novel

486
 487 Table 2: Ablation study on our key components. The numbers in parentheses indicate the specific
 488 rank within the group (lower is better). The 1st, 2nd and 3rd best values are highlighted with
 489 background colors. The **Average Rank** column reflects the overall balanced performance.

Model	Aesthetic Score \uparrow	HPS v2 \uparrow	PickScore \uparrow	ImageReward \uparrow	CLIP Score \uparrow	Average Rank \downarrow
DRSPO(Ours)	5.726 (3)	0.228 (1)	0.134 (2)	0.275 (1)	0.321 (1)	1.6
w/o control	6.009 (1)	0.136 (4)	0.166 (1)	0.171 (3)	0.320 (2)	2.2
w/o DPO	5.734 (2)	0.220 (2)	0.129 (3)	0.247 (2)	0.320 (2)	2.2
w IP-Adapter	5.564 (4)	0.216 (3)	0.086 (4)	-0.058 (4)	0.313 (4)	3.8

490
 491 adaptation of Direct Preference Optimization (DPO) to the MultiDiffusion process, which we apply
 492 at both stages to align the output with human quality preferences. Experimental results validate
 493 our approach, demonstrating a state-of-the-art balance of global structural integrity and fine-grained
 494 details. Future work in long-scroll generation will continue to focus on the central challenge of si-
 495 multaneously achieving coherent global structures and rendering fine-grained, high-fidelity details.

502 7 THE USE OF LARGE LANGUAGE MODELS (LLMs)

503
 504 For the preparation of this manuscript, we utilized Large Language Model (LLM) based tools to
 505 assist with improving the language, clarity, and readability of the prose. We wish to clarify that
 506 the application of these tools was strictly confined to stylistic and grammatical refinement. The
 507 core scientific contributions, including the conceptual framework, methodology, and experimental
 508 analysis, were developed entirely by the authors.

509
 510
 511
 512
 513
 514
 515
 516
 517
 518
 519
 520
 521
 522
 523
 524
 525
 526
 527
 528
 529
 530
 531
 532
 533
 534
 535
 536
 537
 538
 539

540 REFERENCES

541

542 Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for
543 controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023.

544 Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James
545 Kwok, Ping Luo, Huchuan Lu, et al. Pixart- α : Fast training of diffusion transformer for photore-
546 alistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

547

548 Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping
549 Luo, Huchuan Lu, and Zhenguo Li. Pixart- σ : Weak-to-strong training of diffusion transformer for
550 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer,
551 2024.

552 Stanislav Frolov, Brian B Moser, and Andreas Dengel. Spottediffusion: A fast approach for seamless
553 panorama generation over time. *arXiv preprint arXiv:2407.15507*, 2024.

554 Moayed Haji-Ali, Guha Balakrishnan, and Vicente Ordóñez. Elasticdiffusion: Training-free ar-
555 bitrary size image generation through global-local content separation. In *Proceedings of the*
556 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6603–6612, 2024.

557

558 John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaus-
559 sian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal
560 Processing-ICASSP'07*, volume 4, pp. IV–317. IEEE, 2007.

561 Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A
562 reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

563

564 Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in
565 neural information processing systems*, 33:6840–6851, 2020.

566 Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-
567 a-pic: An open dataset of user preferences for text-to-image generation. 2023.

568

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

570

571 Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via
572 synchronized joint diffusions. In *Thirty-seventh Conference on Neural Information Processing
573 Systems*, 2023.

574 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong
575 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to fol-
576 low instructions with human feedback. *Advances in neural information processing systems*, 35:
577 27730–27744, 2022.

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

581

582 Fabio Quattrini, Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara. Merging and splitting diffu-
583 sion paths for semantically coherent panoramas. In *European Conference on Computer Vision*,
584 pp. 234–251. Springer, 2024.

585

586 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea
587 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances
588 in neural information processing systems*, 36:53728–53741, 2023.

589

590 Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi
591 Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An
592 open large-scale dataset for training next generation image-text models. *Advances in neural in-
593 formation processing systems*, 35:25278–25294, 2022.

594

595 Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv
596 preprint arXiv:2010.02502*, 2020.

594 Shoukun Sun, Min Xian, Tiankai Yao, Fei Xu, and Luca Capriotti. Guided and variance-corrected
 595 fusion with one-shot style alignment for large-content image generation. In *Proceedings of the*
 596 *AAAI Conference on Artificial Intelligence*, volume 39, pp. 7114–7121, 2025.

597
 598 Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam,
 599 Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using
 600 direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision*
 601 *and Pattern Recognition*, pp. 8228–8238, 2024.

602 Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind
 603 super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international confer-*
 604 *ence on computer vision*, pp. 1905–1914, 2021.

605 Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li.
 606 Human preference score v2: A solid benchmark for evaluating human preferences of text-to-
 607 image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

608 Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang,
 609 and Jianfei Cai. Taming stable diffusion for text to 360° panorama image generation. In *Proceed-*
 610 *ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

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

615 Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang.
 616 Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of*
 617 *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8018–8027, 2024b.

618 Teng Zhou and Yongchuan Tang. Twindiffusion: Enhancing coherence and efficiency in panoramic
 619 image generation with diffusion models. In *ECAI 2024*, pp. 386–393. IOS Press, 2024.

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648 **A APPENDIX**649 **A.1 INTRODUCTION ON DIRECT PREFERENCE OPTIMIZATION**650 **DPO** The RLHF (Reinforcement Learning from Human Feedback) pipeline comprises three pri-
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
12310
12311
12312
12313
12314
12315
12316
12317
12318
12319
12320
12321
12322
12323
12324
12325
12326
12327
12328
12329
12330
12331
12332
12333
12334
12335
12336
12337
12338
12339
12340
12341
12342
12343
12344
12345
12346
12347
12348
12349
12350
12351
12352
12353
12354
12355
12356
12357
12358
12359
12360
12361
12362
12363
12364
12365
12366
12367
12368
12369
12370
12371
12372
12373
12374
12375
12376
12377
12378
12379
12380
12381
12382
12383
12384
12385
12386
12387
12388
12389
12390
12391
12392
12393
12394
12395
12396
12397
12398
12399
123100
123101
123102
123103
123104
123105
123106
123107
123108
123109
123110
123111
123112
123113
123114
123115
123116
123117
123118
123119
123120
123121
123122
123123
123124
123125
123126
123127
123128
123129
123130
123131
123132
123133
123134
123135
123136
123137
123138
123139
123140
123141
123142
123143
123144
123145
123146
123147
123148
123149
123150
123151
123152
123153
123154
123155
123156
123157
123158
123159
123160
123161
123162
123163
123164
123165
123166
123167
123168
123169
123170
123171
123172
123173
123174
123175
123176
123177
123178
123179
123180
123181
123182
123183
123184
123185
123186
123187
123188
123189
123190
123191
123192
123193
123194
123195
123196
123197
123198
123199
123200
123201
123202
123203
123204
123205
123206
123207
123208
123209
123210
123211
123212
123213
123214
123215
123216
123217
123218
123219
123220
123221
123222
123223
123224
123225
123226
123227
123228
123229
123230
123231
123232
123233
123234
123235
123236
123237
123238
123239
123240
123241
123242
123243
123244
123245
123246
123247
123248
123249
123250
123251
123252
123253
123254
123255
123256
123257
123258
123259
123260
123261
123262
123263
123264
123265
123266
123267
123268
123269
123270
123271
123272
123273
123274
123275
123276
123277
123278
123279
123280
123281
123282
123283
123284
123285
123286
123287
123288
123289
123290
123291
123292
123293
123294
123295
123296
123297
123298
123299
123300
123301
123302
123303
123304
123305
123306
123307
123308
123309
123310
123311
123312
123313
123314
123315
123316
123317
123318
123319
123320
123321
123322
123323
123324
123325
123326
123327
123328
123329
123330
123331
123332
123333
123334
123335
123336
123337
123338
123339
123340
123341
123342
123343
123344
123345
123346
123347
123348
123349
123350
123351
123352
123353
123354
123355
123356
123357
123358
123359
123360
123361
123362
123363
123364
123365
123366
123367
123368
123369
123370
123371
123372
123373
123374
123375
123376
123377
123378
123379
123380
123381
123382
123383
123384
123385
123386
123387
123388
123389
123390
123391
123392
123393
123394
123395
123396
123397
123398
123399
123400
123401
123402
123403
123404
123405
123406
123407
123408
123409
123410
123411
123412
123413
123414
123415
123416
123417
123418
123419
123420
123421
123422
123423
123424
123425
123426
123427
123428
123429
123430
123431
123432
123433
123434
123435
123436
123437
123438
123439
123440
123441
123442
123443
123444
123445
123446
123447
123448
123449
123450
123451
123452
123453
123454
123455
123456
123457
123458
123459
123460
123461
123462
123463
123464
123465
123466
123467
123468
123469
123470
123471
123472
123473
123474
123475
123476
123477
123478
123479
123480
123481
123482
123483
123484
123485
123486
123487
123488
123489
123490
123491
123492
123493
123494
123495
123496
123497
123498
123499
123500
123501
123502
123503
123504
123505
123506
123507
123508
123509
123510
123511
123512
123513
123514
123515
123516
123517
123518
123519
123520
123521
123522
123523
123524
123525
123526
123527
123528
123529
123530
123531
123532
123533
123534
123535
123536
123537
123538
123539
123540
123541
123542
123543
123544
123545
123546
123547
123548
123549
123550
123551
123552
123553
123554
123555
123556
123557
123558
123559
123560
123561
123562
123563
123564
123565
123566
123567
123568
123569
123570
123571
123572
123573
123574
123575
123576
123577
123578
123579
123580
123581
123582
123583
123584
123585
123586
123587
123588
123589
123590
123591
123592
123593
123594
123595
123596
123597
123598
123599
123600
123601
123602
123603
123604
123605
123606
123607
123608
123609
123610
123611
123612
123613
123614
123615
123616
123617
123618
123619
123620
123621
123622
123623
123624
123625
123626
123627
123628
123629
123630
123631
123632
123633
123634
123635
123636
123637
123638
123639
123640
123641
123642
123643
123644
123645
123646
123647
123648
123649
123650
123651
123652
123653
123654
123655
123656
123657
123658
123659
123660
123661
123662
123663
123664
123665
123666
123667
123668
123669
123670
123671
123672
123673
123674
123675
123676
123677
123678
123679
123680
123681
123682
123683
123684
123685
123686
123687
123688
123689
123690
123691
123692
123693
123694
123695
123696
123697
123698
123699
123700
123701
123702
123703
123704
123705
123706
123707
123708
123709
123710
123711
123712
123713
123714
123715
123716
123717
123718
123719
123720
123721
123722
123723
123724
123725
123726
123727
123728
123729
123730
123731
123732
123733
123734
123735
123736
123737
123738
123739
123740
123741
123742
123743
123744
123745
123746
123747
123748
123749
123750
123751
123752
123753
123754
123755
123756
123757
123758
123759
123760
123761
123762
123763
123764
123765
123766
123767
123768
123769
123770
123771
123772
123773
123774
123775
123776
123777
123778
123779
123780
123781
123782
123783
123784
123785
123786
123787
123788
123789
123790
123791
123792
123793
123794
123795
123796
123797
123798
123799
123800
123801
123802
123803
123804
123805
123806
123807
123808
123809
123810
123811
123812
123813
123814
123815
123816
123817
123818
123819
123820
123821
123822
123823
123824
123825
123826
123827
123828
123829
123830
123831
123832
123833
123834
123835
123836
123837
123838
123839
123840
123841
123842
123843
123844
123845
123846
123847
123848
123849
123850
123851
123852
123853
123854
123855
123856
123857
123858
123859
123860
123861
123862
123863
123864
123865
123866
123867
123868
123869
123870
123871
123872
123873
123874
123875
123876
123877
123878
123879
123880
123881
123882
123883
123884
123885
123886
123887
123888
123889
123890
123891
123892
123893
123894
123895
123896
123897
123898
123899
123900
123901
123902
123903
123904
123905
123906
123907
123908
123909
123910
123911
123912
123913
123914
123915
123916
123917
123918
123919
123920
123921
123922
123923
123924
123925
123926
123927
123928
123929
123930
123931
123932
123933
123934
123935
123936
123937
123938
123939
123940
123941
123942
123943
123944
123945
123946
123947
123948
123949
123950
123951
123952
123953
123954
123955
123956
123957
123958
123959
123960
123961
123962
123963
123964
123965
123966
123967
123968
123969
123970
123971
123972
123973
123974
123975
123976
123977
123978
123979
123980
123981
123982
123983
123984
123985
123986
123987
123988
123989
123990
123991
123992
123993
123994
123995
123996
123997
123998
123999
1239999
12399999
123999999
1239999999
12399999999
123999999999
1239999999999
12399999999999
123999999999999
1239999999999999
12399999999999999
123999999999999999
123999999999

702 Subsequently, a lightweight adapter network projects these features to obtain the image conditioning
 703 vector, $c_{\text{image}} = \mathcal{P}(\mathbf{F}(I_{\text{LR}}))$. The final U-Net conditioning is a concatenation of text and image
 704 prompts: $\mathbf{c} = [c_{\text{text}}, c_{\text{image}}]$. This design is highly parameter-efficient and enables flexible, multi-
 705 modal control. Equation 11 is thus modified to:

706

707

708

$$709 \quad I_{t-1}^i = \Phi(F_i(J_t), F_i(I_{\text{canny}}), [\mathbf{c}_{\text{text}}, \mathbf{c}_{\text{image}}]) \quad (33)$$

710

711

712

713 We ablate this block in our ablation study detailed in Sec. 5.1.1, where the results show that incor-
 714 porating the Ip-Adapter block into our method degrades performance. We also find that this block
 715 introduces a blurring effect, which adversely impacts the overall image quality. Moreover, an exces-
 716 sive number of control modules can cause the generated image to adhere too rigidly to the control
 717 blueprint, thereby preventing the achievement of our desired creative outcome.

718

719

720

721

722 A.3 DPO FOR MULTIDIFFUSION

723

724

725 In this section, we will introduce the details of our proposed MultiDiffusion-DPO method. The
 726 base loss function, presented in Equation. 16, is difficult to optimize directly because the reverse
 727 process trajectory $p_{\theta}(x_{1:T}|x_0)$ is intractable. To overcome this, we adopt an approximation strategy
 728 similar to that used in the original DPO paper. We substitute the intractable reverse process with
 729 the tractable forward noising process, $q(x_{1:T}|x_0)$. This substitution, followed by the application of
 730 Jensen’s inequality, allows us to derive a final, optimizable loss function.

731

732

733

734

$$\begin{aligned}
 735 \quad L_{\text{DPO-MultiDiffusion}} &= -\log \sigma \left(\beta \mathbb{E}_{J_{1:T}^w \sim q(J_{1:T}|J_0^w, c), J_{1:T}^l \sim q(J_{1:T}|J_0^l, c)} \right. \\
 736 &\quad \left. \left[\log \frac{p_{\theta}(J_{0:T}^w|c)}{p_{\text{ref}}(J_{0:T}^w|c)} - \log \frac{p_{\theta}(J_{0:T}^l|c)}{p_{\text{ref}}(J_{0:T}^l|c)} \right] \right) \\
 737 &= -\log \sigma \left(\beta \mathbb{E}_{J_{1:T}^w \sim q(J_{1:T}|J_0^w), J_{1:T}^l \sim q(J_{1:T}|J_0^l)} \right. \\
 738 &\quad \left. \left[\sum_{i=1}^T \log \frac{p_{\theta}(J_{t-1}^w|J_t^w)}{p_{\text{ref}}(J_{t-1}^w|J_t^w)} - \log \frac{p_{\theta}(J_{t-1}^l|J_t^l)}{p_{\text{ref}}(J_{t-1}^l|J_t^l)} \right] \right) \\
 739 &= -\log \sigma \left(\beta \mathbb{E}_{J_{1:T}^w \sim q(J_{1:T}|J_0^w), J_{1:T}^l \sim q(J_{1:T}|J_0^l)} T \mathbb{E}_t \right. \\
 740 &\quad \left. \left[\log \frac{p_{\theta}(J_{t-1}^w|J_t^w)}{p_{\text{ref}}(J_{t-1}^w|J_t^w)} - \log \frac{p_{\theta}(J_{t-1}^l|J_t^l)}{p_{\text{ref}}(J_{t-1}^l|J_t^l)} \right] \right) \\
 741 &= -\log \sigma \left(\beta T \mathbb{E}_t \mathbb{E}_{J_{t-1,t}^w \sim q(J_{t-1,t}|J_0^w), J_{t-1,t}^l \sim q(J_{t-1,t}|J_0^l)} \right. \\
 742 &\quad \left. \left[\log \frac{p_{\theta}(J_{t-1}^w|J_t^w)}{p_{\text{ref}}(J_{t-1}^w|J_t^w)} - \log \frac{p_{\theta}(J_{t-1}^l|J_t^l)}{p_{\text{ref}}(J_{t-1}^l|J_t^l)} \right] \right) \\
 743 &= -\log \sigma \left(\beta T \mathbb{E}_{t, J_t^w \sim q(J_t|J_0^w), J_t^l \sim q(J_t|J_0^l)} \right. \\
 744 &\quad \left. \mathbb{E}_{J_{t-1}^w \sim q(J_{t-1}|J_0^w), J_{t-1}^l \sim q(J_{t-1}|J_0^l)} \left[\log \frac{p_{\theta}(J_{t-1}^w|J_t^w)}{p_{\text{ref}}(J_{t-1}^w|J_t^w)} - \log \frac{p_{\theta}(J_{t-1}^l|J_t^l)}{p_{\text{ref}}(J_{t-1}^l|J_t^l)} \right] \right) \quad (34)
 \end{aligned}$$

756 By Jenson’s inequality, it can be applied:
 757

$$\begin{aligned}
 758 \quad & L_{\text{DPO-MultiDiffusion}} \leq \\
 759 \quad & -\mathbb{E}_{t, J_t^w \sim q(J_t | J_0^w), J_t^l \sim q(J_t | J_0^l)} \log \sigma \left(-\beta T \mathbb{E}_{J_{t-1}^w \sim q(J_{t-1} | J_0^w, J_t^w), J_{t-1}^l \sim q(J_{t-1} | J_0^l, J_t^l)} \right. \\
 760 \quad & \left. \left[\log \frac{p_\theta(J_{t-1}^w | J_t^w, c)}{p_{\text{ref}}(J_{t-1}^w | J_t^w, c)} - \log \frac{p_\theta(J_{t-1}^l | J_t^l, c)}{p_{\text{ref}}(J_{t-1}^l | J_t^l, c)} \right] \right) \\
 761 \quad & = -\mathbb{E}_{t, J_t^w \sim q(J_t | J_0^w), J_t^l \sim q(J_t | J_0^l)} \log \sigma \left(-\beta T \left(\right. \right. \\
 762 \quad & \left. \left. + \mathbb{D}_{\text{KL}}(q(J_{t-1}^w | J_0^w, J_t^w) \parallel p_\theta(J_{t-1}^w | J_t^w, c)) \right. \right. \\
 763 \quad & \left. \left. - \mathbb{D}_{\text{KL}}(q(J_{t-1}^w | J_0^w, J_t^w) \parallel p_{\text{ref}}(J_{t-1}^w | J_t^w, c)) \right. \right. \\
 764 \quad & \left. \left. - \mathbb{D}_{\text{KL}}(q(J_{t-1}^l | J_0^l, J_t^l) \parallel p_\theta(J_{t-1}^l | J_t^l, c)) \right. \right. \\
 765 \quad & \left. \left. + \mathbb{D}_{\text{KL}}(q(J_{t-1}^l | J_0^l, J_t^l) \parallel p_{\text{ref}}(J_{t-1}^l | J_t^l, c)) \right) \right) \quad (35) \\
 766 \quad & \\
 767 \quad & \\
 768 \quad & \\
 769 \quad & \\
 770 \quad & \\
 771 \quad & \\
 772 \quad & \\
 773 \quad &
 \end{aligned}$$

774 Thus, the objective function becomes:
 775

$$\begin{aligned}
 776 \quad & L(\theta) \approx -\mathbb{E}_{t, J_t^w \sim q(J_t | J_0^w), J_t^l \sim q(J_t | J_0^l)} \log \sigma \left(-\beta T \left(\right. \right. \\
 777 \quad & \left. \left. \|\mu_t^w - \mu_\theta^w\| - \|\mu_t^w - \mu_{\text{ref}}^w\| - (\|\mu_t^l - \mu_\theta^l\| - \|\mu_t^l - \mu_{\text{ref}}^l\|) \right) \right) \\
 778 \quad & = -\mathbb{E}_{t, J_t^w \sim q(J_t | J_0^w), J_t^l \sim q(J_t | J_0^l)} \log \sigma \left(-\beta T \left(\right. \right. \\
 779 \quad & \left. \left. \|\mu_t^w - \sum_{i=1}^n F_i^{-1}(\mu_{\theta,i}^w)\| - \|\mu_t^w - \sum_{i=1}^n F_i^{-1}(\mu_{\text{ref},i}^w)\| \right. \right. \\
 780 \quad & \left. \left. - (\|\mu_t^l - \sum_{i=1}^n F_i^{-1}(\mu_{\theta,i}^l)\| - \|\mu_t^l - \sum_{i=1}^n F_i^{-1}(\mu_{\text{ref},i}^l)\|) \right) \right) \quad (36) \\
 781 \quad & \\
 782 \quad & \\
 783 \quad & \\
 784 \quad & \\
 785 \quad & \\
 786 \quad & \\
 787 \quad &
 \end{aligned}$$

$$\begin{aligned}
 788 \quad & L(\theta) = -\mathbb{E}_{t, J_t^w \sim q(J_t | J_0^w), J_t^l \sim q(J_t | J_0^l)} \log \sigma \left((-\beta T \omega(\lambda_t)) \left(\right. \right. \\
 789 \quad & \left. \left. \|\epsilon^w - \sum_{i=1}^n F_i^{-1}(\epsilon_\theta(x_t^w, t))\| - \|\epsilon^w - \sum_{i=1}^n F_i^{-1}(\epsilon_{\text{ref}}(x_t^w, t))\| \right. \right. \\
 790 \quad & \left. \left. - (\|\epsilon^l - \sum_{i=1}^n F_i^{-1}(\epsilon_\theta(x_t^l, t))\| - \|\epsilon^l - \sum_{i=1}^n F_i^{-1}(\epsilon_{\text{ref}}(x_t^l, t))\|) \right) \right) \quad (37) \\
 791 \quad & \\
 792 \quad & \\
 793 \quad & \\
 794 \quad & \\
 795 \quad & \\
 796 \quad &
 \end{aligned}$$

797 A.4 MORE RESULTS

798 This section presents further qualitative comparisons against baseline methods. These visual results
 799 highlight our method’s superior ability to maintain global structural coherence while simultaneously
 800 rendering finer and more intricate local details. Furthermore, to elucidate the distinct contributions
 801 of each component, we present qualitative results from our ablation study in Fig. 8. Beyond this
 802 component-level analysis, we also present a diverse portfolio of generated scrolls (Fig. figs. 9 and 10)
 803 to underscore the model’s stylistic flexibility and its capacity for creative generalization.
 804

805 A.5 FAILURE CASES

806 In this section, we analyze several failure cases (Fig.11) to provide a transparent account of our
 807 method’s limitations. First, our method can struggle with generating complex semantic concepts,
 808 such as the human faces shown in Fig.11a. This is a well-documented challenge in text-to-image
 809

810 synthesis and highlights that our framework’s performance is fundamentally dependent on the ca-
 811 pabilities of the underlying base model. Furthermore, while our approach significantly mitigates
 812 object repetition, it does not entirely eliminate this artifact, as instances can still occur during the
 813 generation of very long scrolls (Figs. figs. 11b and 11c). However, we note that the frequency of
 814 such repetitions is substantially reduced compared to baseline methods. Finally, Fig.11d illustrates
 815 a semantic error where, despite a globally coherent structure, the model conflates the concepts of a
 816 “beach” and “water” from the prompt. This suggests that there is still room for improvement in the
 817 model’s text-to-image alignment and its ability to interpret complex spatial relationships.

818

819

A.6 SOCIAL IMPACT

820

821 The primary impact of our work is to advance the technical capabilities within the field of image gen-
 822 eration, specifically for high-aspect-ratio content. Our DRSPO framework addresses the critical and
 823 persistent challenge of maintaining global coherence in long-scroll synthesis, enabling the reliable
 824 creation of large-scale panoramic and narrative images without common artifacts like object repe-
 825 tition. By providing a method that successfully balances global structure with fine-grained detail,
 826 we offer a more practical and effective tool for artists and designers, potentially enhancing creative
 827 workflows in digital art, virtual environment design, and sequential media. Furthermore, our ap-
 828 plication of preference optimization to this task contributes a valuable technique for better aligning
 829 generative outputs with human intent, improving the usability and quality of these powerful creative
 830 tools.

831

832

A.7 REPRODUCIBILITY STATEMENT

833

834 We present a detailed derivation of our method for applying DPO to the MultiDiffusion framework
 835 in Sec.3.2, with an extended analysis provided in AppendixA.3. The hyperparameters and settings
 836 for our training process are outlined in Sec.5.1.2. Furthermore, Sec.5.1.1 describes our dataset
 837 construction pipeline, which includes prompt selection using the Coze platform and the subsequent
 838 generation of preference pairs from various base models. Our implementation will be made publicly
 839 available on GitHub.

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864
865
866
867
868
869
870
871
872



(a) Merge-Attend-Diffuse (MAD)

873
874
875
876
877
878
879
880



(b) MultiDiffusion

881
882
883
884
885
886
887
888



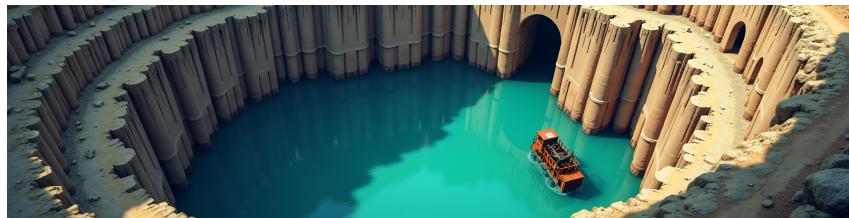
(c) SyncDiffusion

889
890
891
892
893
894
895
896



(d) SDXL

897
898
899
900
901
902
903
904



(e) Flux

905
906
907
908
909
910
911
912



(f) DRSPO(Ours)

913
914
915
916
917

Figure 4: Scrolls(1024×4096) generated using prompts:“Illustrate an old, abandoned quarry. The sheer, tiered rock walls descend to a deep pool of startlingly turquoise water at the bottom. Rusted machinery lies half-submerged, a sign of the industry that once thrived here.”

918
919
920
921
922
923
924
925
926



(a) Merge-Attend-Diffuse (MAD)

927
928
929
930
931
932
933
934



(b) MultiDiffusion

935
936
937
938
939
940
941
942



(c) SyncDiffusion

943
944
945
946
947
948
949
950



(d) SDXL

951
952
953
954
955
956
957
958



(e) Flux

959
960
961
962
963
964
965
966



(f) DRSPO(Ours)

967
968
969
970
971

Figure 5: Scrolls(1024×4096) generated using prompts:“Write about a small, hidden beach where the sand is made of smooth, colorful sea glass. The tiny, worn fragments of glass in green, blue, and brown sparkle in the sun like jewels with every wave that washes ashore.”

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025



(a) Merge-Attend-Diffuse (MAD)



(b) MultiDiffusion



(c) SyncDiffusion



(d) SDXL



(e) Flux



(f) DRSPO(Ours)

Figure 6: Scrolls(1024×4096) generated using prompts:“Paint a picture of a massive radio telescope dish in a remote, quiet valley. The enormous white dish is pointed towards the sky, silently listening for signals from the depths of the universe, a symbol of human curiosity.”

1026
1027
1028
1029
1030
1031
1032
1033
1034
1035



(a) MAD: Aesthetic Score=6.577

1036
1037
1038
1039
1040
1041
1042
1043



(b) DRSPO(Ours): Aesthetic Score=5.870

1044
1045
1046
1047
1048

Figure 7: Although MAD generates a large number of repetitive and illogical objects, it achieves a higher score on the Aesthetic Score. This suggests that the Aesthetic Score may favor images that generate repetitive objects.

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

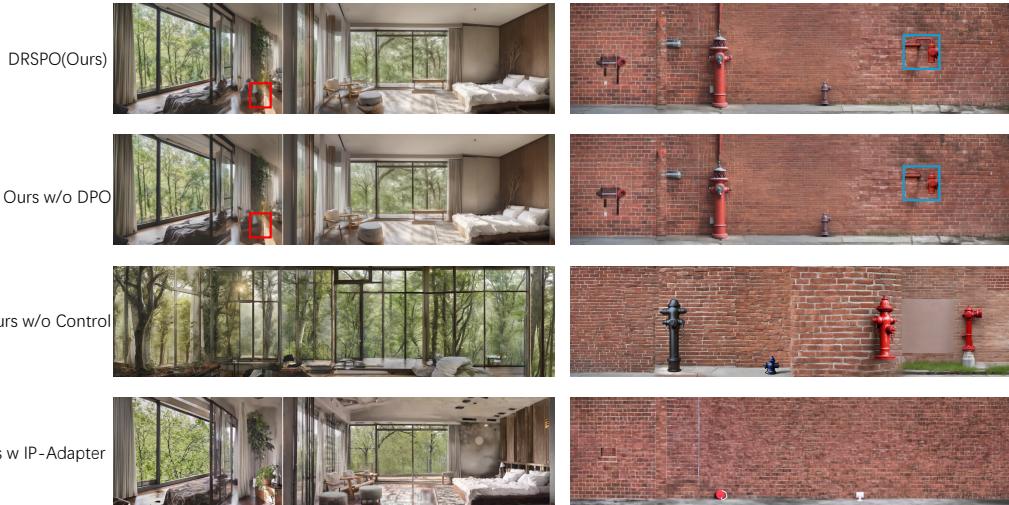
1066

1067

1068

1069

1070



1071
1072
1073
1074
1075
1076
1077
1078
1079

Figure 8: The visual results of our ablation study clearly elucidate the distinct role of each component. Given that the foundational structure is already established by the low-resolution blueprint, the DPO applied in the HR stage is primarily responsible for refining fine-grained details and enhancing local textures. Conversely, the blueprint itself governs the global composition; consequently, its removal leads to a catastrophic collapse of structural integrity, fundamentally altering the entire image. Finally, we observe that integrating an additional IP-Adapter is detrimental, introducing severe visual artifacts such as blurring and the complete omission of objects.

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133



(a) “A lone cyborg ronin standing in a narrow, rain-slicked alley of Neo-Kyoto, steam rising from a ramen stall, the neon glow from flickering holographic advertisements reflecting in the puddles of acid rain, cinematic lighting, volumetric fog, hyperrealistic detail, octane render.”



(b) “A sleek, black aerodyne vehicle hovers silently outside the panoramic window of a megacorporation’s penthouse, overlooking a sprawling, light-saturated metropolis at night, the interior of the vehicle is dark, hinting at a powerful figure within, wide-angle concept art, style of Syd Mead and Blade Runner 2049.”



(c) “A bustling, multi-level bazaar in a grimy undercity, crowded with humans, androids, and aliens bartering for black market cybernetics and glowing bio-enhancements, tangled wires and pipes hang overhead, vibrant chaos, dynamic composition, detailed character design, cinematic atmosphere.”



(d) “An elite netrunner slumped in a customized console chair, wires jacked directly into their neural port, surrounded by a chaotic web of holographic screens displaying cascading lines of cryptic code and abstract data-fortress schematics, the only light source is the glow from the monitors, dark and moody, cyberpunk aesthetic.”

Figure 9: Scrolls generated by our method using cyberpunk style prompts.

1134

1135

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187



(a) “An ancient, colossal tree whose branches hold an entire elven city, with crystalline bridges connecting glowing lantern-lit homes, waterfalls cascading from the highest limbs into a misty abyss below, fantasy concept art, highly detailed, epic scale, style of Studio Ghibli and Ori and the Blind Forest.”



(b) “A lone knight clad in ornate, sun-bleached armor kneels before a dragon’s colossal skeleton in a vast, desolate wasteland of black sand, a glowing magical sword plunged into the ground beside him, twin moons illuminating the scene, cinematic, somber atmosphere, hyperrealistic.”



(c) “A bustling underground dwarven forge deep within a mountain, where rivers of molten gold flow in carved channels and master artisans hammer runes of power onto massive war axes, the air thick with sparks and the heat of the earth’s core, dramatic lighting, dynamic composition, fantasy art.”



(d) “The grand hall of a forgotten king, now sunken beneath the ocean, with massive coral-encrusted thrones and pillars, schools of bioluminescent fish swim through the ethereal, sun-dappled water, a graceful mermaid queen watches from the shadows, underwater photography, magical realism, highly detailed.”

Figure 10: Scrolls generated by our method using fantasy style prompts.

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203



(a) “Write about a crowded open-air market in a foreign city. The air is thick with the smells of spices, street food, and incense, and the sound is a cacophony of vendors shouting, music playing, and people haggling.”



(b) “Describe a vast, multi-level highway interchange, a complex knot of concrete ramps and flyovers. Cars, trucks, and buses flow in a constant, mesmerizing, and intricate dance of organized chaos.”



(c) “Illustrate a vast field of poppies, a sea of brilliant red under a clear blue sky. The delicate petals tremble in a gentle breeze, creating a beautiful and poignant scene that is often associated with remembrance.”



(d) “Depict a vast, empty beach at low tide, stretching for miles. The wet sand reflects the sky like a mirror, and the receding water has left intricate patterns and shallow tidal pools behind.”

Figure 11: we present and analyze several failure cases to provide a transparent account of our method’s current limitations.

1235

1236

1237

1238

1239

1240

1241