SyntheticManga: Training-Free Manga Generation with Phased Diffusion

Xuelei Peng[1] Chi-Keung Tang[1] Yu-Wing Tai[2]
[1]The Hong Kong University of Science and Technology [2]Dartmouth College
xpengat@connect.ust.hk, cktang@cse.ust.hk, yu-wing.tai@dartmouth.edu

Figure 1. **SyntheticManga** achieves a strong balance between character consistency and prompt alignment, producing diverse imagery that follows the artistic and evolving storylines of the input prompt. Notably, consistent characters are not limited to humans but extend to a wide range of entities.

Abstract

Synthesizing visually consistent characters across sequential frames is a fundamental yet largely unsolved challenge in manga generation, where practitioners must navigate a critical trade-off between preserving character identity and faithfully adhering to textual prompts. We introduce ***SyntheticManga****, a training-free framework that reconciles this tension through a principled, phased control strategy over the diffusion sampling trajectory. In the high-noise phase, we propose* ***Boltzmann Fourier Guidance (BFG)****—to our knowledge, the first application of Boltzmann distribution principles to the character-consistency problem—which constructs a probabilistic fusion mask derived from spectral feature drift to adaptively inject structural information from a reference image. In the subsequent mid-noise phase, our* ***Adaptive Drift Modulator (ADM)*** *leverages classical PID control theory to continuously minimize the L_1 drift between noise predictions, thereby enabling fine-grained identity correction. Extensive experiments on the ConsiStory+ benchmark demonstrate that* ***SyntheticManga*** *achieves state-of-the-art performance, attaining a superior balance between identity consistency and prompt fidelity compared to existing methods.*

1. Introduction

Large-scale text-to-image diffusion models have fundamentally transformed digital content creation, catalyzing a surge of interest in automated visual narrative generation—a domain that demands not only high-fidelity image synthesis but also profound coherence and consistency across a sequence of images. Among visual storytelling forms, manga presents a singularly complex challenge that remains largely unaddressed by contemporary generative frameworks: its distinctive artistic style, intricate panel layouts, and deep emphasis on expressive character arcs create relational and sequential demands that far exceed the capabilities of models designed for isolated image generation.

A central obstacle in manga synthesis is maintaining character consistency. State-of-the-art diffusion models frequently fail to preserve a character's identity—encompassing facial features, attire, and overall

appearance—across panels depicting varied poses, expressions, and actions. Conventional methods for enforcing identity confront a difficult *consistency-alignment trade-off*: approaches that rigidly enforce identity stifle prompt-driven actions or emotions, yielding static, lifeless characters, whereas methods that prioritize prompt alignment erode the character's core identity, populating the narrative with seemingly different individuals.

Prior work spans two broad categories. Training-based methods such as DreamBooth [12] and textual inversion [6] fine-tune models to learn a new concept for a specific character; however, they are computationally expensive, require per-subject optimization, and carry the fatal downside of potentially altering the original artistic style of the reference image, often degrading its visual qualities. Training-free methods offer greater efficiency yet introduce distinct trade-offs: StoryDiffusion [18] struggles to maintain identity when the generated panel's aspect ratio deviates from the reference, while concatenation-based approaches such as One-Prompt-One-Story [10] can produce over-similar results and generate rigid, "pasted-on" images, particularly when the prompt specifies large spatial features. Moreover, such approaches suffer from what we term *prompt poisoning*, whereby textual tokens corresponding to one frame's attributes exert undue influence on subsequent frames, compromising textual alignment.

To address these limitations, we propose **SyntheticManga**, a three-phase, training-free framework for layout-aware, multi-character manga generation. Our foundational insight is that robust character consistency is best achieved through *phased control* of the denoising process—front-loading identity anchoring in the early, high-noise stages and then systematically relaxing constraints to permit expressive, prompt-driven detail. The framework comprises two primary novel contributions:

1. **Boltzmann Fourier Guidance (BFG):** Operating in the initial high-noise phase ($t > T_{\text{phase1}}$), BFG imprints the fundamental structural identity of the reference character. To our knowledge, this is the first work to involve the Boltzmann distribution from statistical mechanics in the context of identity-preserving image generation. We formulate a feature drift error E based on the spectral difference between current and reference feature maps; this error informs a probabilistic mask, $P \propto e^{-E/T}$, that adaptively modulates the fusion of frequency components, ensuring strong yet flexible structural resonance.
2. **Adaptive Drift Modulator (ADM):** As another centerpiece of our framework, the ADM operates in the mid-noise phase ($T_{\text{phase2}} < t \leq T_{\text{phase1}}$). We apply the Proportional-Integral-Derivative (PID) control theory directly to the feature space of a diffusion model for identity preservation. The ADM functions as a feedback loop, measuring the L_1 drift between noise predictions of the current and reference latents and applying a continuous, corrective force to the generation trajectory.

In a final refinement phase ($t \leq T_{\text{phase2}}$), all reference injections are strategically zeroed out, allowing the model to synthesize fine-grained details without over-similarity.

Our contributions are threefold: **(i)** a novel, end-to-end training-free framework that jointly addresses layout complexity and robust multi-character consistency; **(ii)** BFG and ADM, which establish a new paradigm for controllable generation grounded in principles from statistical mechanics and control theory; and **(iii)** extensive experiments demonstrating state-of-the-art performance on the *ConsiStory+* benchmark, a finding corroborated by decisive preference in user studies.

2. Related Work

Consistent Generation. Achieving consistent character identity across generated images without costly fine-tuning remains a pivotal challenge in generative modeling. ConsiStory [14] introduces a Subject-Driven Self-Attention (SDSA) block that enables cross-frame attention to subject-specific patches, though its reliance on masking can limit diversity and yield a "pasted-on" appearance. StoryDiffusion [18] is similarly constrained, struggling with multi-character interactions and exhibiting brittleness when panel aspect ratios deviate significantly from the reference, leading to quality degradation. MasaCtrl [3] targets editing rather than narrative variation—a focus ill-suited for the diverse poses and expressions essential in manga. One-Prompt-One-Story [10] exploits the inherent context consistency of text encoders; while clever, it suffers from what we term *prompt poisoning*, where textual tokens corresponding to one frame's attributes impose unintentional influence on subsequent frames, sacrificing textual alignment. It is also prone to generating rigid, pasted-looking images, especially when the prompt specifies large spatial features intended to dominate the generated image.

Training-based approaches such as DreamBooth [12] offer robust identity preservation via subject-specific fine-tuning but are computationally demanding and carry the fatal downside of potentially corrupting the model's learned style priors, often downgrading visual quality and artistic integrity. IP-Adapter [17] injects image features via cross-attention, yet often trades identity preservation for prompt alignment. More recent transformer-based models like FLUX.1 Kontext [1] demonstrate powerful generative capabilities but require pre-training on billions of images, making them inaccessible for broad adaptation. In our experiments, FLUX further exhibits limitations in fine-grained prompt alignment, struggling to render specific compositions such as diverse viewpoints such as side views and bird's eye views. A critical limitation for narrative generation is its tendency to resize inputs to an internal

preferred resolution, thereby disregarding the precise panel layouts essential for manga storytelling. In contrast, **SyntheticManga** is training-free and implements phased control along temporal denoising, embedding physics-inspired mechanisms directly into the generation process. This provides adaptive, phase-wise regulation of both identity and spatial layout—without masking, iterative refinement, or rigid prompt concatenation—enabling higher identity consistency while maintaining the expressive flexibility critical for manga storytelling.

Manga and Layout Generation. Manga synthesis poses unique challenges due to its complex panel layouts and distinct visual grammar. DiffSensei [15] integrates Multimodal Large Language Models (MLLMs) for fine-grained character control but requires specialized adapters and training, and suffers from severe image-quality degradation. MangaDiffusion [4] employs transformer-based blocks for panel coherence yet also relies on a trained model and heavily sacrifices visual quality. Our framework diverges by utilizing the training-free LayoutPrompter [9], which leverages the MangaZero dataset in a retrieval-and-composition manner, granting the ability to generate diverse, narratively coherent page layouts without the computational overhead and architectural modifications of trained layout models, thereby preserving flexibility and efficiency.

Control Theory and Statistical Mechanics in Generative Models. Integrating classical control principles into deep generative models is a sophisticated and emerging frontier. Proportional-Integral-Derivative (PID) controllers [19] are foundational in control systems for minimizing the error between a measured variable and a desired setpoint while ensuring robustness and stability. Prior works such as RCDM [16] and optimal control perspectives on diffusion-based models [2] explore control-theoretic ideas to guide stochastic generation; however, to our knowledge, SyntheticManga's **Adaptive Drift Modulator (ADM)** is the one of the pioneers to apply a PID-like loop directly in the feature space of a diffusion model for identity preservation. By treating character identity as the setpoint and feature-space divergence as the error, the ADM enables adaptive, fine-grained control that corrects past deviations and anticipates future drift, offering a more robust and flexible alternative to static attention mechanisms.

We further draw a novel conceptual parallel from statistical mechanics. The Boltzmann distribution [13] models the probability of a system occupying a state as a function of that state's energy; we posit that our work is the first to operationalize this principle for identity-preserving image generation. Our **Boltzmann Fourier Guidance (BFG)** uniquely re-contextualizes feature-space alignment error as a system's "energy," yielding a physics-inspired, probabilistic mechanism that adaptively modulates identity injection in the frequency domain—establishing it as another centerpiece of our framework's innovation alongside the ADM.

3. Method

We present our training-free manga generation framework, which combines layout generation with a multi-phase denoising process to ensure prompt-adherent character consistency. Input prompts for each manga panel are obtained either from raw user inputs or LLM-generated results. We employ LayoutPrompter [9], a training-free retrieval-and-composition module, to generate page layouts from a story description. Leveraging the MangaZero dataset [15], LayoutPrompter analyzes narrative requirements such as character count and scene transitions, retrieves analogous layouts, and composes new page structures. The output provides precise panel coordinates along with bounding boxes for characters and dialogue, serving as a structural guide for subsequent image generation. This training-free approach enables diverse, conventional manga layouts without costly fine-tuning or architectural modification.

The core of our contribution to identity preservation is a multi-phase generation process that dynamically manages the influence of a reference character image throughout denoising (Figure 2), ensuring robust identity replication in the early, formative phases while allowing prompt-driven flexibility and refinement in later phases.

3.1. Phase 1: Boltzmann Fourier Guidance (BFG)

Motivation and Theoretical Grounding. In the initial high-noise timesteps ($t > T_{\text{phase1}}$), the latent tensor z_t is largely stochastic and lacks coherent structure. The primary objective is to imprint the fundamental visual identity of the reference character without rigidly overwriting the nascent structure, which would stifle prompt-driven variation. This requires guidance that is strong when generated and reference structures align yet gracefully attenuates as they diverge, preventing artifact introduction.

We model reference-feature fusion as a probabilistic process by drawing from the mathematical structure of the Boltzmann distribution, which provides a principled way to define a probability distribution over a set of states based on their "energy." In our context, a "state" is the configuration of the current feature map, and we define its "energy" $E(t)$ as a metric of its structural dissimilarity from the reference feature map. A low-energy state corresponds to high structural alignment, making it a high-probability state that should be strongly reinforced. The exponential form of the Boltzmann distribution ($P \propto e^{-E/kT}$) is chosen for its specific mathematical properties, which are ideally suited to this problem:

1. Adaptive, Non-Linear Scaling. It provides a non-linear mapping where guidance strength is high (approaching 1) for low-dissimilarity states but decays smoothly and rapidly as dissimilarity increases. This ensures that only

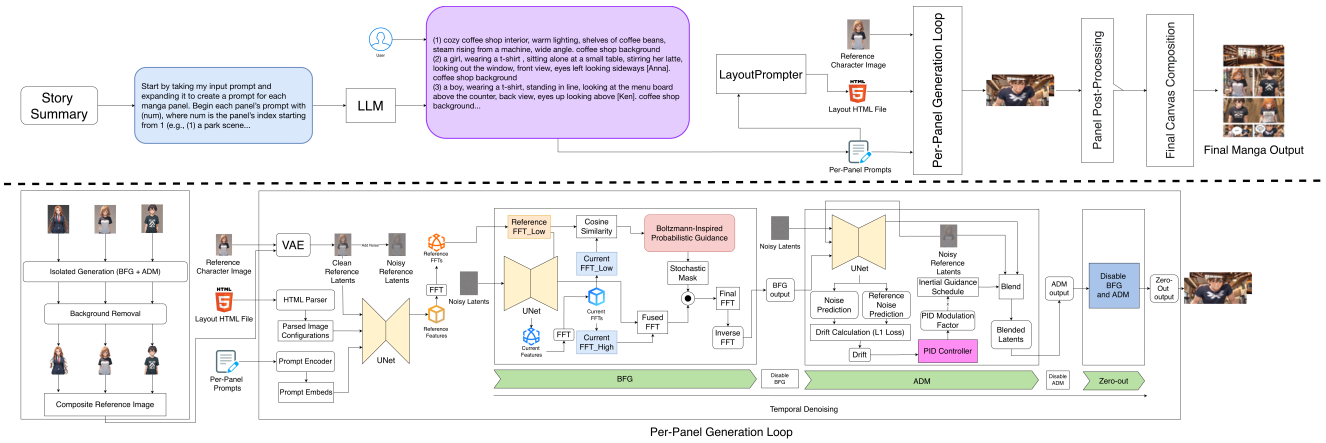

Figure 2. **The pipeline of SyntheticManga.** Top shows the overall pipeline; bottom is a detailed illustration of the per-panel generation loop. In the top figure, the process in the box labeled "Panel Post-Processing" adds dialog bubbles onto the generated manga panels. In the bottom figure, the blue box represents the "zero-out" operation. Please zoom in for details.

well-aligned reference features are fused, while poorly-aligned features, which would likely cause artifacts, are suppressed.

2. Principled Control. The temperature hyperparameter T provides a theoretically grounded "knob" to control sensitivity: a low T creates a stricter filter accepting only states with very high alignment, while a higher T allows softer, more permissive fusion.

This probabilistic framework provides a robust theoretical foundation for adaptively guiding the early stages of diffusion, moving beyond heuristic or empirically-tuned blending factors.

Formulation. The BFG process unfolds in three steps at each timestep t in Phase 1:

1. Feature Drift as Energy. We define the drift error $E(t)$ by first transforming the intermediate feature maps of the current latent, F_{curr}, and the reference latent, F_{ref}, into the frequency domain via a 2D Fast Fourier Transform (FFT). The error is then computed as one minus the cosine similarity between the magnitudes of their respective low-frequency components:

$$E(t) = 1 - \cos\_\text{sim}(|\mathcal{F}_{\text{low}}(F_{\text{curr}})|, |\mathcal{F}_{\text{low}}(F_{\text{ref}})|)$$

where $\mathcal{F}_{\text{low}}$ denotes the low-pass filtered Fourier spectrum. This metric effectively quantifies the structural dissimilarity.

2. Stochastic Mask Generation. The calculated error $E(t)$ informs a probabilistic scaling mask, $M_{\text{stochastic}}$, for each channel of the feature map, governed by a temperature hyperparameter T:

$$M_{\text{stochastic}}(t) = \exp\left(-\frac{E(t)}{T}\right)$$

The temperature T controls the mask's sensitivity to the error, analogous to the "temperature" parameter in the original formulation of the Boltzmann distribution.

3. Spectral Fusion and Modulation. We hypothesize that an effective guidance signal can be formed by fusing the stable, low-frequency (structural) components of the reference spectrum with the noisy, high-frequency (detail) components of the current spectrum. This fused spectrum, $\mathcal{F}_{\text{fused}}$, is formulated as:

$$\mathcal{F}_{\text{fused}} = (\mathcal{F}(F_{\text{curr}}) \odot M_{\text{high-pass}}) + (\mathcal{F}(F_{\text{ref}}) \odot M_{\text{low-pass}})$$

This fused spectrum is then probabilistically applied via element-wise multiplication with $M_{\text{stochastic}}$, which serves to validate the fusion hypothesis: larger mask values exaggerate the fusion, while lower values discourage it. The result is transformed back into the spatial domain via an inverse FFT to yield the guided feature map, which replaces the original in the UNet's forward pass.

3.2. Phase 2: Adaptive Drift Modulator (ADM)

Motivation and Theoretical Justification. The text prompt, via Classifier-Free Guidance (CFG), acts as a powerful disturbance force that constantly pushes the generation trajectory away from the reference character's appearance, causing an "identity drift".. While the CFG force steers the trajectory towards the target prompt manifold (e.g., "a character reading a book"), our goal is to simultaneously constrain it to the identity manifold represented by the reference character; the persistent CFG disturbance threatens drift from this manifold, resulting in a loss of character consistency. This frames our task as a classical problem of trajectory regulation and disturbance rejection in a dynamical system, for which the Proportional-Integral-Derivative (PID) controller is a foundational and theoretically optimal solution. The choice of a full PID controller is specifically motivated by the need to address three distinct temporal characteristics of the error:

1. Proportional (P) Term. Provides an immediate corrective force proportional to the current drift.

2. Integral (I) Term. Eliminates steady-state error. A persistent conflict between the prompt and the reference can create a small but constant drift that a P-controller alone cannot nullify. The I-term accumulates this past error over time, amplifying the corrective action until the persistent drift is eliminated.

3. Derivative (D) Term. Provides anticipatory control and system damping. By responding to the error's rate of change, it anticipates future drift and smooths the corrective action, preventing the overshooting and oscillations that can manifest as visual artifacts when the guidance is too aggressive.

Formulation. The ADM operates as a closed-loop feedback system at each timestep t in Phase 2:

1. Drift Calculation. We precisely quantify the feature-space drift $d(t)$ using two noise predictions from the UNet ϵ_θ. The first, ϵ_{target}, is the standard prediction for the current latent z_t conditioned on the target prompt C_{target}. For the second, we construct a noisy reference latent $z_{\text{ref},t}$ by adding the same amount of noise present in z_t to the clean reference latent $z_{\text{ref},0}$, then compute a parallel noise prediction ϵ_{ref} for $z_{\text{ref},t}$ conditioned on a simple reference prompt C_{ref}. The drift is defined as the L_1 loss between these two noise predictions:

$$d(t) = ||\epsilon_\theta(z_t, t, C_{\text{target}}) - \epsilon_\theta(z_{\text{ref},t}, t, C_{\text{ref}})||_1$$

2. PID Control. This error signal $d(t)$ is fed into the discrete-time PID controller to compute a modulation factor $M_{\text{ADM}}(t)$:

$$P(t) = K_p \cdot d(t)$$
$$I(t) = I(t-1) + K_i \cdot d(t)$$
$$D(t) = K_d \cdot (d(t) - d(t-1))$$
$$M_{\text{ADM}}(t) = 1.0 + P(t) + I(t) + D(t)$$

where K_p, K_i, and K_d are the proportional, integral, and derivative gains, respectively.

3. Modulated Latent Blending. The modulation factor $M_{\text{ADM}}(t)$ scales a predefined, linear base guidance schedule $\mathcal{G}(t)$ to produce a final blending factor

$$\alpha_{\text{final}}(t) = \mathcal{G}(t) \cdot M_{\text{ADM}}(t)$$

This factor directly guides the current latent z_t by blending it with the noisy reference latent $z_{\text{ref},t}$ before the scheduler step:

$$z'_t = (1 - \alpha_{\text{final}}(t)) \cdot z_t + \alpha_{\text{final}}(t) \cdot z_{\text{ref},t}$$

The standard denoising step is then performed using the original noise prediction ϵ_{target} applied to the corrected latent z'_t, creating a subtle but powerful corrective nudge at each step that steers the generation trajectory back towards the reference identity manifold.

3.3. Phase 3: Refinement and Zero-Out

In the final low-noise timesteps ($t \leq T_{\text{phase2}}$), we perform a strategic **zero-out** of all reference injections. This is not merely a cessation of guidance but an essential transfer of control. The robustness of the identity anchoring performed by BFG and ADM in the earlier, high-impact phases makes continued intervention unnecessary and counterproductive. By liberating the model from the rigid constraints of the reference features, this phase allows the UNet to synthesize fine-grained, high-frequency details dictated purely by the text prompt, mitigating over-similarity and ensuring characters integrate naturally into the scene.

3.4. Multi-Character Panel Synthesis

A common failure mode in multi-subject generation is "character fusion," where distinct identities bleed into one another. We address this with a robust three-stage workflow. First, each character specified in the prompt is generated individually at high resolution using our full three-phase pipeline, ensuring each possesses a strong, well-preserved identity that serves as a reference for the final composite. Second, their backgrounds are removed and the resulting foregrounds are composited onto a single transparent canvas according to the panel's layout coordinates. Third, this composite canvas is used as the reference for a final generation pass that employs the full panel prompt containing all characters, executed with intensified identity-preservation hyperparameters to generate a new, prompt-aligned background while keeping the character appearances intact.

3.5. Hyperparameter Rationale

The choice of hyperparameters differs between the single-character and the more demanding multi-character workflow. A longer duration for the identity-preserving phases (achieved with smaller phase-ending timestep numbers) leads to higher reference similarity. In the multi-character workflow, the reference image is the carefully constructed composite of already well-preserved characters, and the primary goal is to maintain their integrity with maximum fidelity while generating a new, coherent background. Accordingly, we employ stronger identity-preserving hyperparameters: the BFG and ADM phases are extended and controller gains are intensified, ensuring robust identity separation and preservation against the strong, potentially conflicting influence of the complex background prompt.

4. Experiments

4.1. Experimental Setups

Comparison with SOTA Methods. We compare our method against a suite of state-of-the-art training-free consistent generation approaches—FLUX.1 Kontext, One-Prompt-One-Story, StoryDiffusion, DiffSensei, and

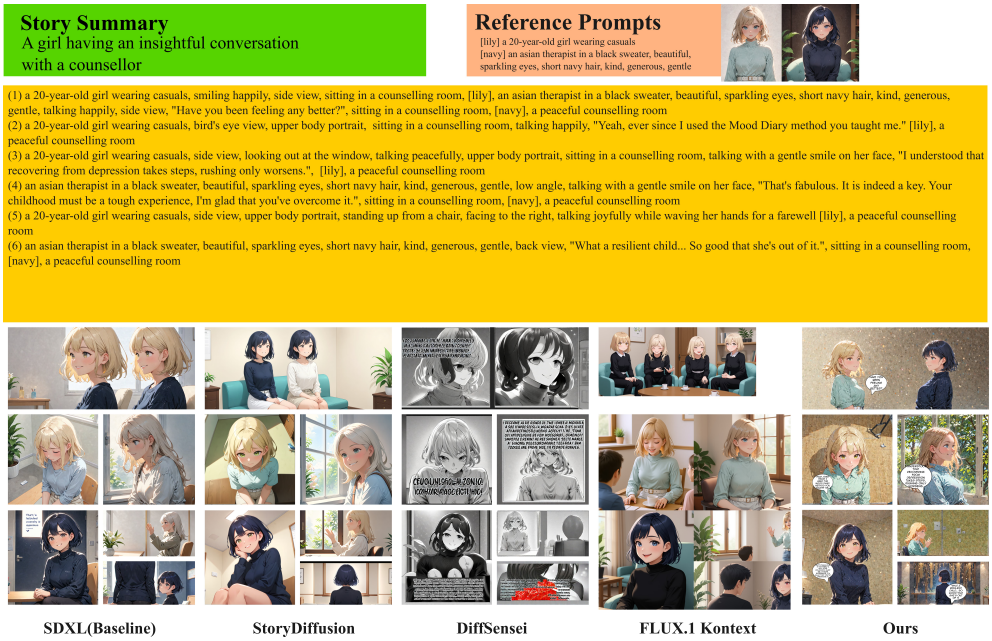

Figure 3. **Qualitative Comparison.** Existing methods either enforce multiple characters to share the same appearance or fail to respect non-human entities (e.g. clothes), leading to identity collapse or missing elements. In contrast, our framework maintains distinct character identities, captures diverse poses and expressions, and preserves compositional and aesthetic coherence across all elements in the scene.

ConsiStory—as well as the base SDXL model as a performance baseline. To ensure a comprehensive and challenging evaluation, we utilize 1000 prompts from the extensive ConsiStory+ benchmark, as introduced by Liu et al. [10].

Evaluation Metrics. We employ CLIP-T [7] for prompt alignment and DreamSim [5] for identity consistency, which correlates strongly with human perceptual judgment. To quantify visual quality, we also report the Fréchet Inception Distance (FID) [8].

4.2. Experimental Results

Qualitative Comparison. Qualitative comparisons in Figure 3 and Figure 6 illustrate the practical advantages of our framework. Existing approaches often enforce multiple characters to share the same appearance or fail to respect non-human entities such as clothing, props, or non-human characters, resulting in identity collapse or missing elements. In contrast, our framework maintains distinct character identities, captures a richer diversity of poses and expressions vital for narrative progression, and preserves compositional and aesthetic coherence across all scene elements. Furthermore, it demonstrably produces images with superior aesthetic appeal, balanced panel composition, and coherent spatial relationships between characters and objects, ensuring that each generated scene faithfully represents individual character traits while maintaining the integrity of interactions and narrative context.

Quantitative Comparison. Figure 4 shows that our method achieves state-of-the-art performance in balancing identity preservation, textual alignment, and visual quality, outperforming all compared approaches. Points closer to the upper-right corner indicate a better balance between identity preservation and prompt alignment; FID scores are labeled on each method, and ours is closest to the upper-right corner with the best FID.

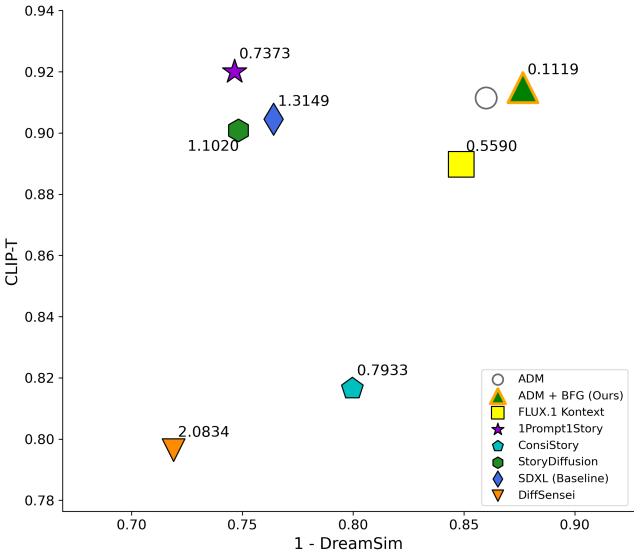

Figure 4. Trade-off between identity preservation and prompt alignment. Points near the upper-right indicate better balance; our method achieves the best overall balance with the lowest FID, outperforming other approaches.

User Study. To assess alignment with human perceptual

Table 1. User Study: Average Scores.

Criteria	SDXL	StoryDiffusion	DiffSensei	FLUX.1 Kontext	Ours
Identity Preservation	3.2101	3.1176	2.3697	3.8559	**3.9076**
Prompt Alignment	2.8824	2.8571	2.3445	2.8644	**3.8235**
Image Quality	3.4622	3.2773	2.4538	3.4915	**3.8151**
Storytelling Ability	2.7563	2.6218	2.5462	2.9068	**3.9160**
Total Score	12.3110	11.8738	9.7142	13.1186	**15.4622**

judgment, we conducted a user study comparing SyntheticManga against SDXL, DiffSensei, FLUX.1 Kontext, and StoryDiffusion. Twenty-four users evaluated each framework based on a holistic assessment of identity preservation, prompt alignment, image quality, and storytelling ability. Table 1 tabulates the findings, revealing a decisive preference for SyntheticManga and confirming its substantially better quality and alignment with human creative intent.

Ablation Study. Figure 4 and Figure 5 present our ablation study validating the efficacy of each component. The sole inclusion of the ADM yields a significantly stronger DreamSim than both the baseline SDXL and every other compared model, demonstrating the ADM's strength even when the BFG component is disregarded. Adding the BFG together with the "zero-out" operation drastically improves identity preservation and prompt alignment, dropping DreamSim and boosting CLIP-T.

Reference Image ADM Only ADM + BFG

Figure 5. **Qualitative Ablation.** The reference prompt is: "a happy girl, pink eyes, wearing a jacket and trousers". The modification prompt is: "reading a book, eyes down, dutch angle". The identity preservation is strongest when ADM and BFG are both activated, while only activating ADM produces inconsistencies e.g. the pink jacket is more inconsistent-looking.

5. Conclusion

We introduced **SyntheticManga**, a training-free, multi-phased framework that reconciles character consistency with prompt alignment in manga generation by establishing a new paradigm grounded in two physics-inspired mechanisms: **Boltzmann Fourier Guidance (BFG)**, which leverages Boltzmann distribution principles for robust structural imprinting, and the **Adaptive Drift Modulator (ADM)**, which applies PID control theory to dynamically correct feature drift. This phased approach, culminating in a strategic zero-out of guidance, secures character identity in the early diffusion stages while preserving creative flexibility for prompt-driven details. Extensive experiments confirm state-of-the-art performance, outperforming existing methods on the ConsiStory+ benchmark and in qualitative user studies, demonstrating the significant potential of applying principles from control theory and statistical mechanics to generative AI as a robust and scalable solution for narrative synthesis.

References

[1] Black Forest Labs (2025). Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2, 10

[2] Julius Berner, Lorenz Richter, and Karen Ullrich. An optimal control perspective on diffusion-based generative modeling. *arXiv preprint arXiv:2211.01364*, 2022. 3

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023. 2

[4] Siyu Chen, Dengjie Li, Zenghao Bao, Yao Zhou, Lingfeng Tan, Yujie Zhong, and Zheng Zhao. Manga generation via layout-controllable diffusion. *arXiv preprint arXiv:2412.19303*, 2024. 3

[5] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 6

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patash-

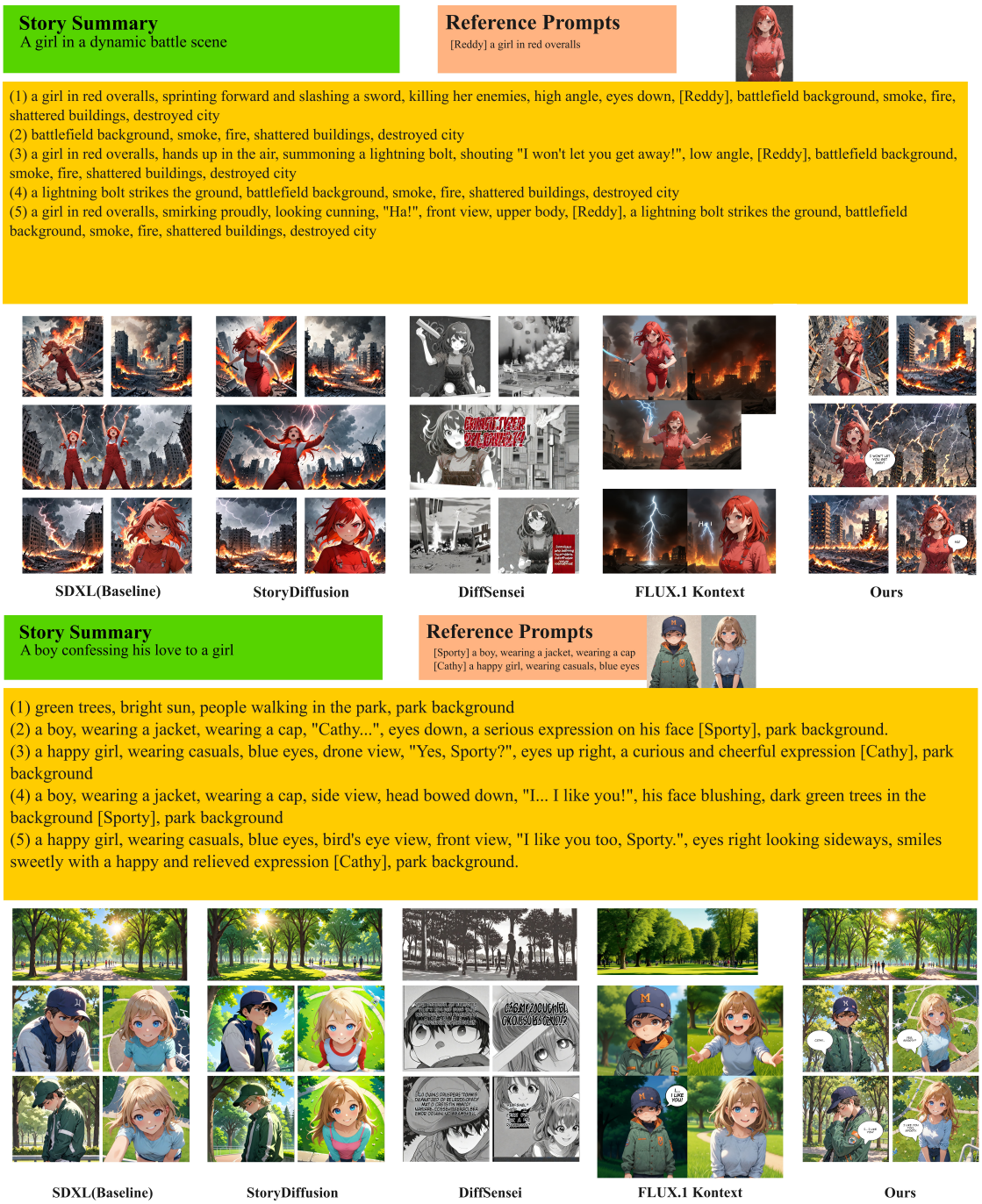

Figure 6. **Additional Qualitative Comparison.** The results presented in this figure compare the outputs of our framework with those from prominent baselines. Please zoom in for clarity.

nik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2022. 6

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2018. 6

[9] Jiawei Lin, Jiaqi Guo, Shizhao Sun, Zijiang James Yang, Jian-Guang Lou, and Dongmei Zhang. Layoutprompter: Awaken the design ability of large language models. *arXiv preprint arXiv:2311.06495*, 2023. 3

[10] Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang,

and Ming-Ming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. *arXiv preprint arXiv:2501.13554*, 2025. 2, 6, 10

[11] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 10

[12] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2023. 2

[13] Kim Sharp and Franz Matschinsky. Translation of ludwig boltzmann's paper "on the relationship between the second fundamental theorem of the mechanical theory of heat and probability calculations regarding the conditions for thermal equilibrium" sitzungberichte der kaiserlichen akademie der wissenschaften. mathematisch-naturwissen classe. abt. ii, lxxvi 1877, pp 373-435 (wien. ber. 1877, 76:373-435). reprinted in wiss. abhandlungen, vol. ii, reprint 42, p. 164-223, barth, leipzig, 1909. *Entropy*, 17(4):1971–2009, 2015. 3

[14] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *arXiv preprint arXiv:2402.03286*, 2024. 2, 10

[15] Jianzong Wu, Chao Tang, Jingbo Wang, Yanhong Zeng, Xiangtai Li, and Yunhai Tong. Diffsensei: Bridging multi-modal llms and diffusion models for customized manga generation. *arXiv preprint arXiv:2412.07589*, 2024. 3, 10

[16] Weifeng Xu, Xiang Zhu, and Xiaoyong Li. Rcdm: Enabling robustness for conditional diffusion model. *arXiv preprint arXiv:2408.02710*, 2024. 3

[17] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2

[18] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 2, 10

[19] Karl Åström and Tore Hägglund. *PID Controllers, 2nd Edition E*. 1995. 3