

Figure 7. **More Additional Qualitative Results.** Additional results from our method.

Appendix

A. Additional Qualitative Results

Further results in Figure 7 and Figure 8 confirm the exceptional performance of our framework in generating complex, multi-character manga pages with high textual alignment and aesthetic appeal. Our method demonstrates a remarkable capacity to preserve the intrinsic qualities of the base model—a crucial attribute that quantitative metrics often fail to capture—affirming its state-of-the-art standing in producing coherent, high-quality narrative visuals.

A.1. Qualitative Results with Realistic Styles

To demonstrate the versatility of our framework beyond manga, we conducted qualitative evaluations on consistent character generation in realistic styles. As illustrated in Fig-

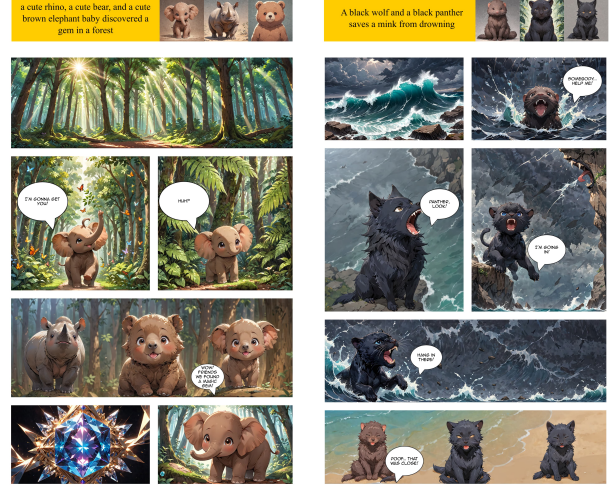


Figure 8. **Additional Qualitative Results On Three Consistent Characters.** Additional results showing our method capable of generating more than two consistent characters within the same manga panel.

ure 9, our method successfully maintains the identity of photorealistic subjects, demonstrating that SyntheticManga functions as a versatile control paradigm for narrative consistency applicable to a wide range of diffusion base models.

B. Additional Qualitative Comparison Results

Qualitative comparisons in Figure 6 and Figure 10 underscore the practical superiority of our method by revealing significant limitations in current state-of-the-art approaches. The baseline SDXL [11] offers no mechanism for character consistency; StoryDiffusion [18] exhibits poor reference similarity and fails to capture the dynamic storytelling nature of manga; DiffSensei [15] struggles with weak character similarity, producing low-quality, monochrome outputs; and FLUX.1 Kontext [1] does not support multi-character panel generation and lacks prompt consistency, especially when tricky camera angles are involved. In stark contrast, our framework delivers an outstanding balance between robust reference similarity and precise prompt alignment, resulting in aesthetically superior images that cohere into compelling narratives. For a fair and direct qualitative comparison, ConsiStory [14] and One-Prompt-One-Story [10] have been omitted from this evaluation, as their architectures do not natively support image-based conditioning.

B.1. Prompt Inconsistency, Incapable Multi-character Generation, and Altered Image Style in FLUX.1 Kontext

As shown in Figure 11, FLUX.1 Kontext is particularly incapable in generating prompt-aligned camera angles and

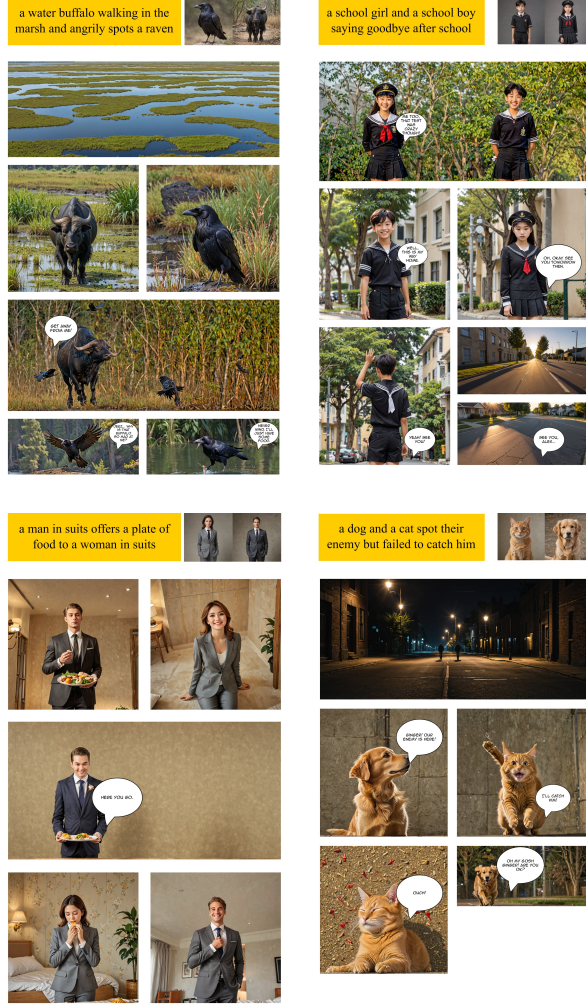


Figure 9. **More Additional Qualitative Results with Realistic Styles.** Additional results with photorealistic styles from our method.

high-quality, non-fused multi-character images, and fails to preserve the anime style by producing realistic-looking images—which we deduce is due to its massive training dataset consisting of mostly realistic imagery. In contrast, our method performs significantly better in these three critical areas and is adaptable to any base model style.

C. Additional Ablation Study

To empirically validate the distinct contribution of each component within our phased framework, we conducted a comprehensive qualitative additive ablation study, with results visualized in Figure 12, confirming that each phase serves an indispensable, synergistic role.

We further investigated the sensitivity of the temperature parameter T within the Boltzmann-inspired BFG, as

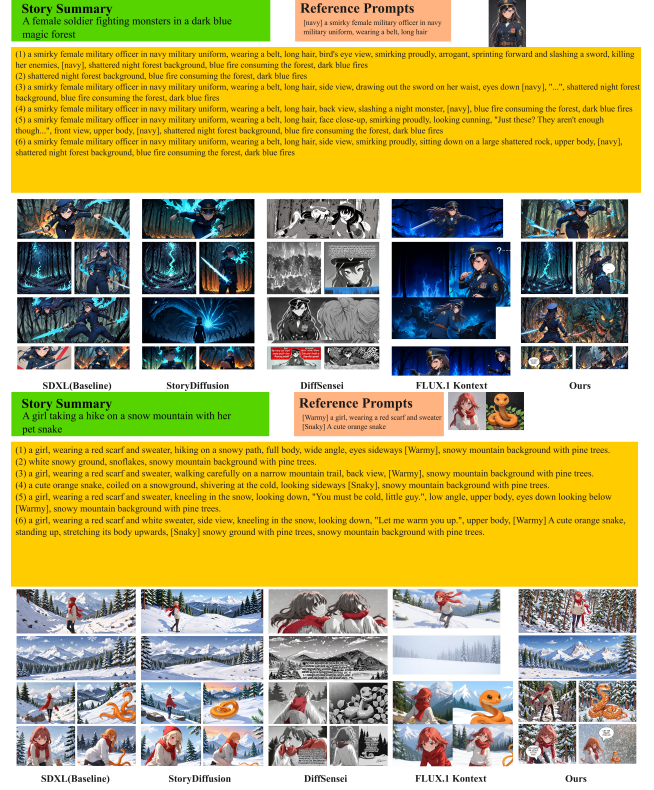


Figure 10. **More Additional Qualitative Comparison.** Additional results showcasing the difference between our method with SDXL, StoryDiffusion, DiffSensei and FLUX.1 Kontext.

illustrated in Figure 13. Lower values ($T < 0.4$) sharpen the probability distribution, approximating a hard selection mask that frequently introduces high-frequency artifacts during the spatial feature establishment stage suitable for low-frequency structures, thus producing reference inconsistency. Conversely, higher values ($T > 0.6$) flatten the distribution towards uniform blending, diluting the adaptive structural guidance and weakening identity preservation. An intermediate value of $T = 0.4$ strikes the optimal balance, maintaining robust structural resonance while ensuring smooth, artifact-free feature integration.

While other weighting functions are possible, our Boltzmann formulation is the most compatible with diffusion sampling and maximum-entropy principles. By treating feature drift as an energy $E(t)$, Boltzmann weighting ($e^{-E/T}$) aligns naturally with the Langevin-style dynamics underlying diffusion. Boltzmann’s exponential form more effectively suppresses high-energy (misaligned) features, resulting in more stable, training-free identity preservation. To empirically validate the necessity of the BFG component, we benchmarked it against two simpler fusion baselines, as illustrated in Figure 14: Linear Interpolation (LERP), which uniformly blends the frequency spec-



Figure 11. **Comparison with FLUX.1 Kontext.** FLUX.1 Kontext often struggles with prompt-aligned camera angles, incapability in generating multi-character panels, and can alter a reference image’s original artistic style.

tra of the reference and current features, and Hard Masking, which uses a fixed similarity threshold to deterministically switch between spectra. Both methods produced reference inconsistency, in stark contrast with our method. BFG’s Boltzmann-inspired probabilistic modulation facilitates a principled and smooth fusion, strongly reinforcing identity where features align while gracefully attenuating guidance where they diverge, confirming that its adaptive nature is crucial for avoiding the characteristic artifacts of more simplistic fusion techniques.

Lastly, an empirical additive ablation in Figure 15 validates the necessity of the full PID architecture for the Adaptive Drift Modulator. The P-only controller suffers from much higher reference dissimilarity than our full ADM method while still exhibiting identity preservation effects; the PI controller improves reference similarity drastically but remains less robust than the complete configuration. Only the full PID simultaneously provides the best and most reliable identity preservation.

D. Implementation Details

Single-Character Configuration. For standard panels featuring a single subject, the BFG is active during the high-noise interval $t \in (1000, 850]$ with temperature $T = 0.4$. Subsequently, the ADM operates in $t \in (850, 750]$ with PID gains $K_p = 1.5$, $K_i = 2.0$, and $K_d = 1.5$. The final zero-out phase commences at $t < 750$.

Multi-Character Configuration. The multi-character workflow relies on a pre-composited reference image containing already identity-consistent character instances; consequently, the primary objective shifts to preserving this composite with extreme fidelity against the prompt’s influence. To achieve this, the identity-enforcing phases are extended and control gains intensified: the BFG phase spans $t \in (1000, 650]$, the ADM phase operates for $t \in (650, 550]$, and the PID gains are increased to $K_p = 2.5$, $K_i = 3.5$, and $K_d = 3.5$. This stronger, prolonged intervention ensures that the complex spatial arrangement and individual identities of the composite reference are maintained throughout background generation.

E. User Study Details

We conducted a user study with 24 participants. As illustrated in Figure 16, for each task, participants were presented with a complete one-page manga narrative—including the high-level story summary, reference characters, and detailed per-panel prompts—alongside several full-page manga generations from our method and competing baselines. They were asked to select the single best entry based on a holistic evaluation of four criteria: Reference Similarity, Textual Consistency, Image Quality, and Story-telling Ability. This protocol was designed to capture the nuanced, multifaceted qualities of a compelling visual narrative that are often missed by quantitative analysis.

F. Discussion

Our work demonstrates that phased, training-free control of the diffusion trajectory can achieve state-of-the-art identity preservation while maintaining high prompt alignment. The framework’s performance is governed by well-defined hyperparameters; through systematic sweeps, we determined optimal configurations that are highly effective across a wide range of scenarios, attesting to the robustness of the phased design. The modular nature of the BFG and ADM phases further allows intuitive fine-tuning should a specific narrative context require it.

From a practical standpoint, SyntheticManga offers a significant usability advantage over monolithic, full-page generation frameworks, where a minor modification to a single panel’s prompt necessitates regenerating the entire page, often yielding unpredictable and inconsistent results



Figure 12. **Additional Additive Ablation Study.** The image in the first row is the reference. Images generated by ADM-only produces more reference inconsistencies, such as inconsistent attire in the first generated image, and the more dissimilar eye and hair colors, while our fully enabled method produces better identity preservation.

across unchanged panels. In contrast, our per-panel generation process ensures that artists and creators can iteratively refine individual story moments with surgical precision, modifying one panel’s prompt without affecting the layout or content of its neighbors—a workflow far more aligned with real-world creative practice.

The multi-pass workflow for robust multi-character synthesis introduces a slight computational overhead compared to single-pass approaches. Quantitatively, our framework records an average execution time of 49.23 ± 8.99 seconds and an average peak VRAM usage of 18.39 ± 0.91 GB, compared to the baseline SDXL’s 27.33 ± 6.92 seconds and 15.54 ± 0.40 GB. We posit that this increase is a justifiable engineering trade-off: the capability to generate consistent, high-quality multi-character manga panels represents a significant functional advancement that outweighs the additional latency, particularly given that comparable baselines fail to resolve identity conflation in complex multi-subject scenarios and our method can still generate a full-page manga in under one minute. Furthermore, our current implementation consists of unoptimized research code; more compute resources are clearly called for by the character consistency and prompt fidelity demands of our phased

diffusion process. Future work could explore distilling the consistency learned during multi-pass generation into a more lightweight, single-pass model.

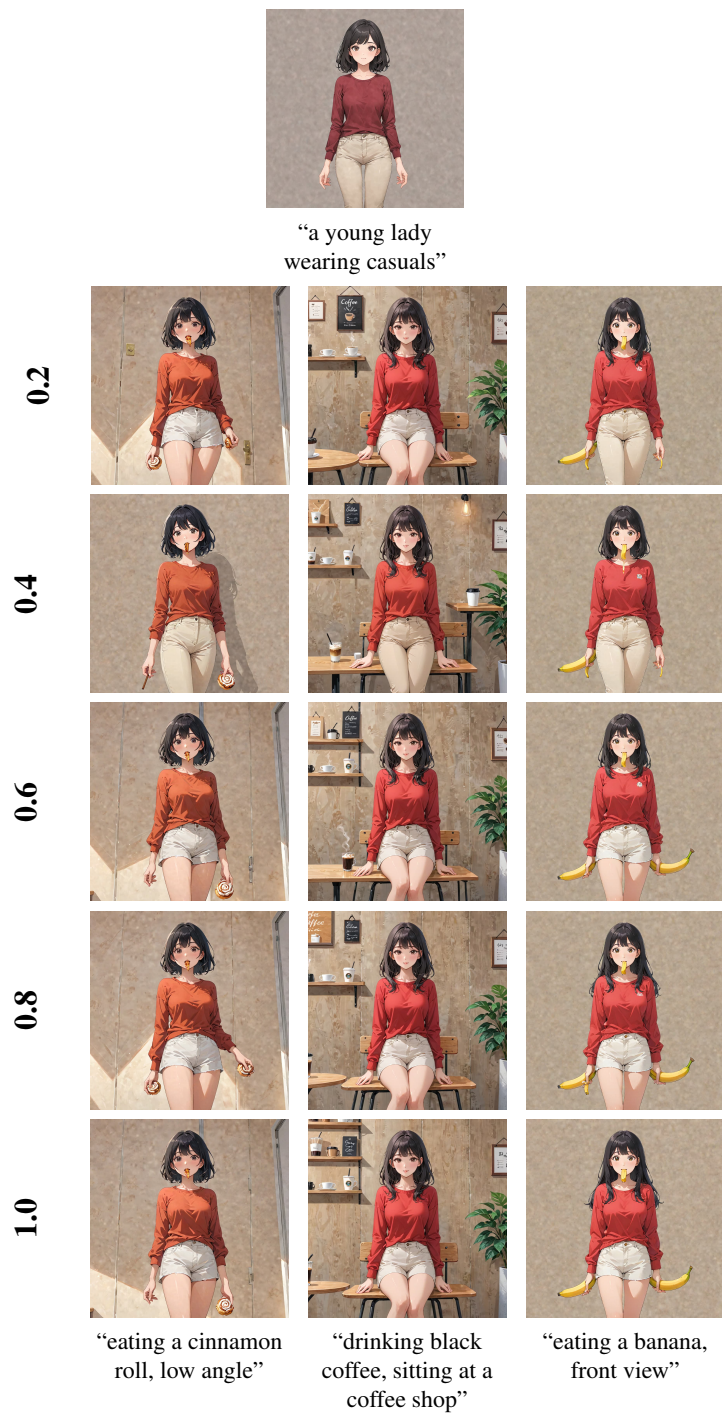


Figure 13. **Ablation of the Temperature Hyperparameter.** The image in the first row is the reference. Varying the temperature hyperparameter demonstrates that an intermediate value ($T \approx 0.4$) is necessary to balance the enforcement of structural identity and expressive diversity.

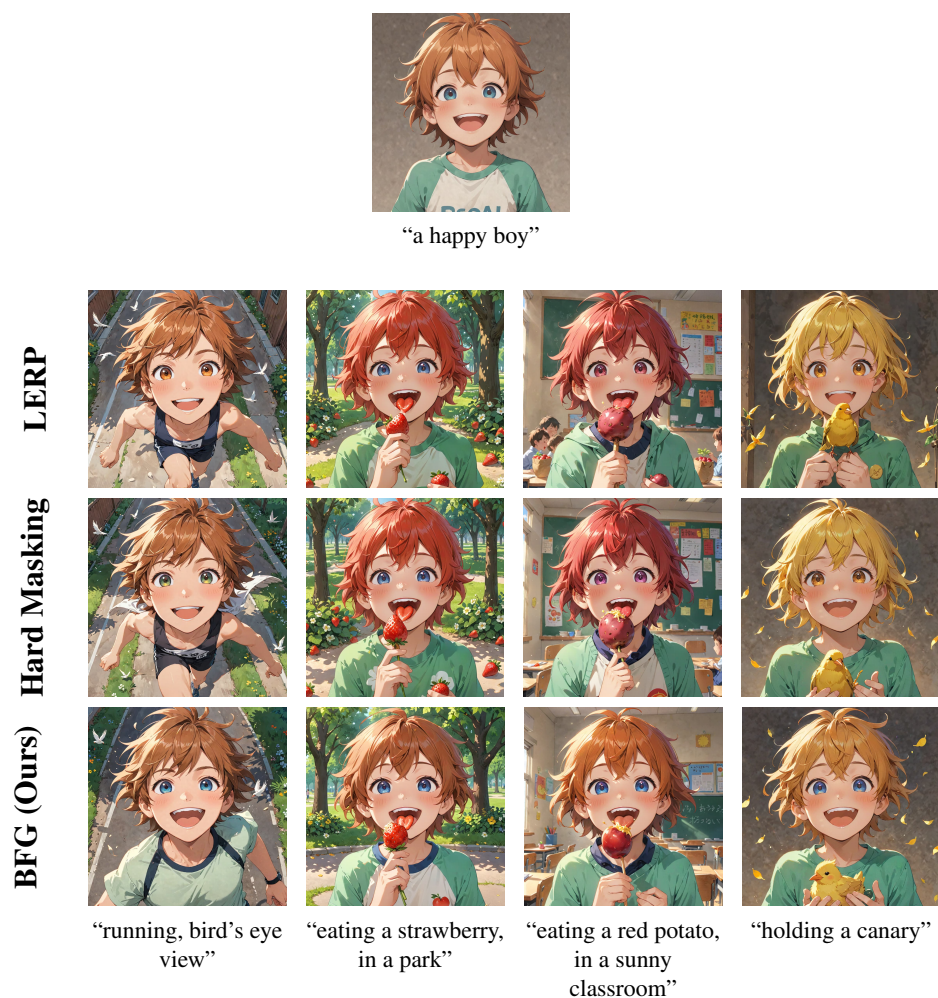


Figure 14. **Ablation of Variants of Boltzmann Fourier Guidance (BFG).** The image in the first row is the reference. Our proposed Boltzmann-derived method (BFG) uniquely preserves shared and prompt-driven features, outperforming its variants Linear Interpolation (LERP) and Hard Masking, which introduce reference dissimilarity particularly in hair colors and attire.



Figure 15. **Ablation of Adaptive Drift Modulator (ADM).** The image in the first row is the reference. Empirical results demonstrate that the full PID architecture is essential for robust control, eliminating the reference inconsistency observed in P-only and PI-Only.

Manga Generation User Study

Guidance Before You Start - IMPORTANT

1. Goal: Select the Top Manga
- You will be presented with a series of one-page manga sets. Your task is to evaluate each one based on these areas:
- 1. Reference Similarity: How similar are the characters in the manga to the reference images
 - 2. Textual Consistency: How well it aligns with the text descriptions in the yellow box
 - 3. Image Quality: how aesthetically detailed the images are
 - 4. Storytelling Ability: how well the story is told, how immersed you feel while reading it
2. Enter in the order of "Reference Similarity", "Textual Consistency", "Image Quality" and "Storytelling Ability" for FIVE score-sets, each score in the score-set is from **1 to 5**, while 5 stands for the highest and best score, 1 stands for the lowest and worst score. Enter for each of the four manga **from left to right**, separated by comma.
3. Example for **ONE question**: 1452, 3441, 5321, 5431, 5455. This means the 1st manga gets the score "1452", meaning its Reference Similarity = 1, Textual Consistency = 4, Image Quality = 5, Storytelling Ability = 2; the same principle for the 2nd, 3rd, 4th and 5th manga. **Please see this example for EACH question during answering.**

1. Enter FIVE score-sets separated by comma based on the four criteria. Example for THIS single question: 1452, 3441, 5321, 5431, 5455. This means the 1st manga gets the score "1452", meaning its Reference Similarity = 1, Textual Consistency = 4, Image Quality = 5, Storytelling Ability = 2; the same principle for the 2nd, 3rd, 4th and 5th manga.

- Your task is to evaluate each one and choose the single best entry based on these areas:
- Reference Similarity: How similar are the characters in the manga to the reference images
 - Textual Consistency: How well it aligns with the text descriptions in the yellow box
 - Image Quality: how aesthetically detailed the images are
 - Storytelling Ability: how well the story is told, how immersed you feel while reading it

Important Note: the 4th manga have a layout that's different and more messy than other ones, please feel free to take into account this messy layout in the evaluation.

Story Summary

A boy confessing his love to a girl

Reference Prompts

[Image of a boy] [Image of a girl]

[Prompt 1: A boy, wearing a jacket, wearing a cap, "Cathy", eyes down, a serious expression on his face [Sporty], park background.]

[Prompt 2: A happy girl, wearing casuals, blue eyes, drone view, "Yes, Sporty!", eyes up right, a curious and cheerful expression [Cathy], park background.]

[Prompt 3: A boy, wearing a jacket, wearing a cap, side view, head bowed down, "I... I like you!", his face blushing, dark green trees in the background [Sporty], park background.]

[Prompt 4: A happy girl, wearing casuals, blue eyes, bird's eye view, front view, "I like you too, Sporty!", eyes right looking sideways, smiles sweetly with a happy and relieved expression [Cathy], park background.]

SIXXL (Rawlinec)

StoryDiffusion

DiffPencil

FLUX.1 Kontext

Ours

Figure 16. **User Study Details.** Participants were given clear instructions and criteria for selection.