Narrative Weaver: Towards Controllable Long-Range Visual Consistency with Multi-Modal Conditioning

Supplementary Material

This supplementary document provides additional details, visualizations, and analyses to support the claims and experiments presented in the main paper. Beyond expanding the empirical evidence, we also further clarify the methodological significance of Narrative Weaver.

Narrative Weaver is not merely an integration of existing components, but a systematic methodology designed to address the challenge of Long-Range Visual Consistency. Specifically, it bridges the gap between short-form visual generation and professional production workflows by enabling high-level narrative logic to consistently govern low-level visual details. Moreover, it establishes a transferable framework for long-range consistency: core components such as the *Dynamic Memory Bank* and *Dual-path Alignment* are modular and can be extended to related domains, including long-form video generation. Finally, we introduce a practical multi-stage progressive training strategy with customized attention mechanisms that efficiently decouple and align complex features, demonstrating strong empirical effectiveness even under resource constraints.

The remainder of this supplementary document is organized as follows:

- **Additional Details on Data Construction (Sec. 7)** provides a deep dive into our novel data creation methodology. We detail the multi-step prompt engineering pipeline and describe the curation process for our new dataset.
- **Evaluation Details (Sec. 8)** outlines the specifics of our evaluation framework. We detail the curation of our test sets, the precise implementation and setup for all baseline methods to ensure fair comparisons, and the methodology of our human evaluation study.
- **Experimental Details (Sec. 9)** is dedicated to the implementation and training of Narrative Weaver. We present the complete training recipe for our multi-stage strategy, including a detailed breakdown of all hyperparameters to ensure full reproducibility.
- **Additional Experimental Results (Sec. 10)** presents an extensive gallery of additional qualitative results. This includes more visual examples from Narrative Weaver and numerous side-by-side comparisons against baselines to further substantiate our claims of superior consistency and aesthetic quality.
- **Detailed Ablation Results (Sec. 11)** provides a quantitative analysis of the contribution of key components within our architecture. These detailed ablation studies validate our design choices and demonstrate the importance of each module.
- **Additional Efficiency Analysis (Sec. 12)** presents a quantitative comparison of the computational cost of our proposed architecture against a vanilla self-attention baseline.
- **Limitations and Future Works (Sec. 13)** discusses the current limitations of our work and outlines promising directions for future research, including the extension to video generation and the need for broader dataset creation.

We believe these supplementary details will provide a comprehensive understanding of our work and its contributions.

Template for Prompt Generation

Role:

As a professional AI advertising design assistant, your task is to conceptualize and generate a set of advertising [Shot Descriptions] with strong storytelling and marketing appeal, based on the product information I provide.
I will provide you with the following information:
1. Product Name: {item_title}
2. Core Selling Point: {selling_point}
3. Recommended Copy/Slogan: {caption}

Please strictly adhere to the following requirements when generating these [Shot Descriptions] :
1. Task: Generate [Shot Descriptions] that satisfy the above information.
2. Product Emphasis: In each [Shot Description], you must clearly and naturally highlight and emphasize the product, making it the visual focus or a key element. Enlarged close-up shots can be used, but avoid showing text like LOGOs or slogans.
3. Marketing Intent: The [Shot Descriptions] should showcase the product creating positive experiences in different scenarios.
4. Style Requirements: Realistic, with high-quality cinematic feel (Cinematic). Use expressive lighting and shadows, with rich and believable details.
5. Detailed Description: For each shot, provide an extremely detailed description that will serve directly as guidance for image generation. The description should include but not be limited to:

- Characters: Appearance, clothing, age, expression, emotional state.
- Product: Can include the product's appearance, material, details, etc.
- Scene: Specific environmental details, props, background elements.
- Action: The character's specific behavior, how the product is used or interacted with.
- Shot Guidance: Shot size (e.g., close-up, medium shot, full shot), camera angle (e.g., eye-level, low angle, high angle), potential camera movement (e.g., dolly in/out, pan, tilt – although the image is static, the description should convey a sense of dynamics).
- Lighting & Color: Specific lighting type (e.g., natural light, studio light, backlight), direction, intensity, as well as the overall color palette and atmosphere.
- Mood & Atmosphere: Convey the intended emotion or atmosphere of the shot (e.g., joyful, relaxed, surprised, professional).

6. Very, very, very strictly follow the output format below for each shot's description and shot guidance. Do not include any content outside the format or extra text. Please ensure consistency and clarity of the format.

Output Format:

Detailed Description: [Please fill in all the detailed descriptions here, including product, characters, scene, action, mood, lighting, color palette, props, etc. Example: A woman around 25 years old leans sideways against a floor-to-ceiling window. The morning sun shines through the sheer curtains onto her slightly wavy chestnut brown hair. She wears a beige turtleneck, holding a hot latte with both hands; the semi-turtleneck fits her neckline, and the lace trim glows softly in the backlight. Outside the window is a street covered in fresh snow, with the warm yellow indoor light contrasting the cool tones outside.]
Shot Guidance: [Please fill in the shot size, angle, and movement requirements for the first shot here. Example: Medium shot, eye-level slightly high angle, lens dollying in slightly towards the character.]

Figure 7. The prompt for Qwen3-30B-A3B to generate text instructions for reference image generation.

7. Additional Details on Data Construction

This section elaborates on the prompt generation methodology outlined in the main text. Our image generation pipeline involves three distinct LLM calls to sequentially produce: (1) shot descriptions for reference images, (2) scenario recommendations for subsequent keyframes, and (3) detailed captions for keyframe synthesis. We provide carefully engineered prompt templates in this section (Fig. 7, Fig. 8, Fig. 9), which are being released to support community research. Our workflow is divided into two main stages:

Stage 1: Reference Image Generation. Our process begins with a dataset of 33,000 real-world product listings, encompassing product names, selling points, and recommendations across categories such as apparel, accessories, home goods, and bags. This information is fed into our self-deployed Qwen3-30B-A3B model [71]. Following the template in Fig. 7, the model generates a detailed textual description for the reference image. To align with e-commerce requirements, these prompts are specifically engineered to convey marketing intent and include rich details covering the subject, product, action, and lighting. This detailed prompt is then used with Qwen-Image [64] to synthesize the final reference image.

Template for Scenario Recommendations

Role:

You are a fashion-focused visual AI. Given the product image (a person modeling clothes) and product metadata, analyze the clothing style and recommend environment-only lifestyle scenes suitable for advertising this product.
Product meta
- Item: {item}
- Selling point: {selling_point}
- Caption: {caption}

Output Format:

- Return ONLY valid JSON (no markdown fences, no extra text).
- JSON schema:

```
{{
  "product_name": "< from 'Item', <= 20 words, no identity/appearance>",
  "suggested_scenes": [
    "<environment-only setting 1>",
    "<environment-only setting 2>",
    "<environment-only setting 3>",
    "<environment-only setting 4>",
    "<environment-only setting 5>",
    "<environment-only setting 6>"
  ]
}}
```

Rules:

- "product_name": product name (max 20 words). Do NOT mention gender, age, face, body, or accessories.
- "suggested_scenes": each entry must describe only the environment/setting/mood/activity area (e.g., café terrace, riverside promenade, bookstore corner, playground area, gallery lobby).
- Do NOT mention a person, identity, gender, age, body, or clothing. Avoid words like "wearing", "boy", "girl", "child", "man", "woman".
- Keep scenes realistic and varied; tie subtly to Selling point/Caption if useful.
- Language: English only.
- CRITICAL: DO NOT use placeholder text, repetitive underscores ('_'), excessive whitespace, or nonsensical strings. Every word in 'product_name' and 'suggested_scenes' must be meaningful and descriptive.
- Return ONLY the JSON object.

Figure 8. Prompt Template for Qwen3-30B-A3B in Generating Keyframe Scenario Recommendations.

Stage 2: Keyframe Synthesis. With the reference image and original product information, we proceed to generate subsequent keyframes. We employ the commercial model Flux.1-Kontext [4] for this task, chosen for its strong capability to maintain strict consistency with a reference image. The prompt generation for this stage is a two-step process:

Scenario Recommendation: First, using the product information and the reference image, we query our Qwen3-30B-A3B model with the template shown in Fig. 8. The model suggests a series of scenes that are contextually appropriate for the product.

Editing Instruction Generation: Next, these recommended scenarios are transformed into precise editing instructions for Flux.1-Kontext using the template in Fig. 9. The specific formatting in this template is crucial for ensuring the instructions are correctly interpreted by the generation model. Notably, we include the string `IMG_1018.CR2` in the prompt, a technique we empirically found to enhance output image quality.

Through extensive experimentation, we summarize several critical insights:

Template for Flux.1-kontext Generation

Role:

You are a senior advertising scene designer.
Imagine you're designing storyboard scenes for a premium lifestyle commercial. Based on the product details, generate 8 rich, cinematic, photographic prompts focused ONLY on the background, setting, mood, and lighting and action. Do NOT describe clothing or the person.
Picture Description: {description_text}

Rules:

Each prompt must:

- Start with: IMG_1018.CR2 This person is...
- Be photorealistic and richly descriptive of the environment, the person's action, and the atmosphere.
- Output strictly as 8 separate lines, each starting with "IMG_1018.CR2 This person is...".
- Each sentence MUST have at least 40 English words and no more than 45. After writing each sentence, count the number of words. If it's less than 40, revise and expand with more ambient, emotional, or environmental context until it reaches the minimum.
- The person must interact with the environment in a physical, intentional way.
- All 8 scenes must be believable real-life moments that match the *type of product* and its lifestyle use.
- Focus on believable urban lifestyle moments with elegant visual mood.

Creative guidance:

- Make the person's action vivid, physical, and specific (e.g., sketching, photographing, flipping pages, gazing into the distance, arranging flowers). Avoid vague or passive verbs like "standing" or "looking."
- Ensure the person physically interacts with the environment in a believable way.
- Avoid abstract emotional labels (e.g., "feeling peaceful").
- Vary sentence rhythm and structure to avoid repetition, while maintaining cinematic flow.
- Make it feel like a high-end, live-action scene from a TV series or fashion editorial.

Tone & Style:

- Use cinematic, naturalistic language.
- Focus only on environmental elements, mood, ambient light, cinematic camera angle, and the person's action and emotions.
- Include implied camera angles or cinematic shot types (e.g., wide shot through a window, overhead view, silhouette at sunset) to make each prompt feel like a film still.

Restrictions:

- Do NOT mention any clothing, accessories, or physical appearance.
- Do NOT describe the person's identity, gender, or race.
- Do NOT use brackets, parentheses, quotation marks, or bullets.

Examples:

- IMG_1018.CR2 This person is gently arranging a bouquet of wildflowers in a rustic vase atop a weathered wooden table, with golden light streaming through a latticed window, casting delicate shadows.
- IMG_1018.CR2 This person is walking through a historic park at dusk, holding a glowing lantern that softly illuminates the cobbled path while long shadows stretch into the deepening blue.
- IMG_1018.CR2 This person is seated on a moss-covered stone step near a quiet forest spring, skimming fingers across the water's surface as dragonflies flit through the shimmering light.

Figure 9. The prompt for Qwen3-30B-A3B to generate text instructions for Flux.1-kontext image generation.

- **Isolated Scenario Planning:** Separating scenario recommendation as an independent subtask proves essential, as inappropriate settings lead to unrealistic outputs resembling mere copy-paste effects.
- **Reference Image Criteria:** To ensure successful subsequent editing, reference images should preferably feature full-body frontal poses of single subjects, avoiding multiple entities or reflective surfaces.
- **Instruction-Oriented Keyframe Prompts:** Keyframe generation benefits from imperative-style descriptions that explicitly guide content creation.
- **Consistency Preservation:** Maintaining cross-frame consistency requires avoiding detailed character and apparel specifications, instead emphasizing scene interactions and background elements.
- **Input Sensitivity of the Generation Model:** We observed that Flux.1-Kontext is highly sensitive to the reference image. For optimal results, the input must be a single-subject, frontal-view photograph with a natural expression. Reference images containing multiple individuals or unconventional poses consistently lead to generation failures.
- **Subject-Agnostic Prompting for Consistency:** To preserve the subject's facial identity and apparel, it is crucial to avoid describing these attributes in the keyframe prompts. Instead, we refer to the subject generically (e.g., starting the prompt with "This person...") and focus exclusively on describing the new scene, action, or camera angle. This prevents the model from regenerating the subject's appearance, thereby ensuring cross-frame consistency.

Data Filtering. Despite our meticulous prompt engineering, a subset of generated images may still fail to meet the stringent quality and consistency standards required for e-commerce. Consequently, we introduce a final data filtering stage. This process addresses two main issues: (1) common-sense violations, such as a subject wearing short sleeves in a snowy winter scene, and (2) common generative artifacts, particularly anatomical inconsistencies like unnatural limbs. This automated filtering is carried out by the Qwen2.5-VL-32B [2] model to ensure the final dataset's high quality.

Why EAVSD? Existing datasets are not well suited for long-range narrative generation. *CoMM* suffers from limited visual quality, *CI-VID* contains only short sequences (typically fewer than three shots per instance), and *OmniGen2* exhibits substantial textual redundancy that limits narrative diversity. These limitations restrict their applicability to professional long-range visual storytelling tasks.

In contrast, EAVSD is specifically designed to support such scenarios. It provides (1) **high-quality** visuals with rich professional annotations, (2) **long-range** sequences with an average of more than eight shots per instance, and (3) structured **narrative logic** aligned with professional production workflows.

As detailed in this section, we have already demonstrated the quality of our data construction pipeline, including model selection, prompt design, and filtering strategies.

Dataset Visualization. Our constructed dataset currently comprises 36K samples, totaling approximately 330K high-quality images, and we plan to expand it with additional product categories in the future. To demonstrate its quality and diversity, we conclude this section with a visual showcase of our final dataset (Fig. 10). The displayed examples cover a wide array of the existing categories, including men's, women's, and children's apparel, loungewear, footwear, and accessories.

Figure 10. **Sample sequences from our EAVSD dataset.** The figure showcases the dataset's diversity across multiple e-commerce categories. Each row displays a sequence where a consistent subject and product are placed in various scenes, guided by descriptive text prompts. The dataset is designed to train models on tasks requiring high visual consistency while allowing for controlled narrative changes in action and setting, which is critical for advertising applications.

8. Evaluation Details

This section provides further details on our dataset processing and evaluation metrics for the experiment in Sec. 5.

8.1. Consistent Visual Generation (Q1)

Test Set Curation. We observed that the original OmniGen2 [65] pre-training data contains a significant number of low-quality samples. To establish a more reliable benchmark, we manually curated a test set of 100 sequences, each comprising alternating text prompts and video frames. Our curation process specifically prioritized samples that demand strong inter-frame consistency, thereby enabling a focused evaluation of Narrative Weaver's capabilities.

LLM-based Evaluation. We provide the prompt template used for our LLM-based evaluation in Fig. 11. This prompt is meticulously designed to comprehensively assess the model's performance across three key dimensions: instruction following, consistency preservation, and image quality. To ensure the validity and reliability of the automated scoring, we instruct the language model to provide a detailed rationale for each assigned score. This practice significantly enhances the stability and trustworthiness of the evaluation process.

Baseline Implementation Details. For a fair comparison, we reproduced several baseline methods. In the following, we describe their implementation details.

- **StoryDiffusion [78]:** This method first generates an initial image from the text prompt corresponding to the reference image. It then utilizes the intermediate tokens from this initial generation process to condition the creation of all subsequent images.
- **AnimeShooter [49]:** The original implementation of AnimeShooter trains a specific LoRA module for each film or IP to achieve high fidelity. To evaluate its generalization capabilities in a broader context, we omitted this LoRA module in our experiments.
- **Reference-based Methods (IP-Adapter [73], Flux.1-Kontext [4], Qwen-Image-Edit [64]):** This category of methods, including IP-Adapter, Flux.1-Kontext, and Qwen-Image-Edit, conditions the generation of each new image on both the initial reference image and the current text prompt. While this approach effectively preserves consistency between each generated image and the initial reference, it often struggles to maintain consistency among the generated images themselves.

Unless a model was specifically trained at a fixed resolution, all baselines were configured to generate images at the same resolution as the provided condition image.

User Study Details. For our human evaluation, we compared Narrative Weaver with the three best-performing methods from Q1 (Flux.1-Kontext [4], Qwen-Image-Edit [64], and StoryDiffusion [78]). Each survey presented participants with the outputs from these four methods for a randomly selected test case. The order of the results was randomized to prevent bias. Participants were asked to choose the most preferable result overall. The final results were compiled from over 180 valid survey responses.

Template for Consistency Evaluation

Text-Image Consistency:

You are a professional image and text evaluator.
Please evaluate the consistency between each generated image and its corresponding text description. Focus on how accurately the description reflects the visual content of its paired image. Please provide an overall score (1-10) and explanation for all pairs.

Your response must be a JSON object:

```
{
    "score": <int_score>,
    "explanation": <string_explanation>
}
```

The following are the image-text pairs for evaluation:

Generation-Condition Consistency:

You are a professional image and text evaluator.
Please compare the reference (input) image with the generated images. Evaluate the consistency of the generated images with the reference image in terms of style and content. Please provide an overall score (1-10) and explanation for all generated images.

Your response must be a JSON object:

```
{
    "score": <int_score>,
    "explanation": <string_explanation>
}
```

The following is the reference image, followed by the generated images:

Generation-Style Consistency:

You are a professional image and text evaluator.
You will receive a series of generated images. Please evaluate the visual style consistency among these images. Please provide a score (1-10) and explanation.

Your response must be a JSON object:

```
{
    "score": <int_score>,
    "explanation": <string_explanation>
}
```

The following are the generated images for evaluation:

Generation-Content Consistency:

You are a professional image and text evaluator.
You will receive a series of generated images. Please evaluate the content consistency among these images (e.g., objects, scenes, overall narrative progression, if applicable). Please provide a score (1-10) and explanation.

Your response must be a JSON object:

```
{
    "score": <int_score>,
    "explanation": <string_explanation>
}
```

The following are the generated images for evaluation:

Image-Quality:

You are a professional image and text evaluator.
You will be receiving a series of generated images. Please evaluate the overall visual quality of these images (e.g., sharpness, realism, artifacts). Provide an overall score (1-10) and explanation for all images.

Your response must be a JSON object:

```
{
    "score": <int_score>,
    "explanation": <string_explanation>
}
```

Below are the generated images for evaluation:

Figure 11. The prompt template provided to GPT-4o for our consistency evaluation. It requires the model to score instruction following, consistency, and image quality, and to provide a rationale for each score.

8.2. Autonomous Narrative Planning (Q2)

CoMM Dataset Processing. To evaluate Narrative Weaver's autonomous narrative generation capability (Q2

in Sec. 5), we employ the CoMM dataset [10]. The original dataset is compiled from diverse sources and suffers from significant data imbalance due to invalid URLs. We therefore selected two instruction-based subsets, Instructables and WikiHow, which are particularly well-suited for assessing the model's problem-solving and narrative planning capacity. For a fair comparison, we re-evaluated the official test set using the weights provided by the original benchmark authors.

During data processing, we standardized each sample by limiting it to a maximum of 16 images and the "`step_info`" field to 12 elements. For continuation tasks, we generated training samples by randomly truncating text-image sequences, using the first half as input and the second half as the target output, ensuring each target contained at least one image. After filtering for invalid images, this procedure yielded approximately 170K training samples. For question-based response tasks, a similar filtering process resulted in approximately 150K training samples. In this experiment, all images were rescaled to a resolution of 512×512 pixels to maintain consistency with the original benchmark.

A Note on the EMU2 Baseline. The official CoMM benchmark includes EMU2 [55], a 33B large-scale unified model. However, we excluded it from our comparison for two primary reasons. First, the official repository does not provide the specific checkpoint or training code used for the benchmark, hindering reproducibility. Second, our preliminary tests with the publicly available EMU2 weights revealed significant failure modes: the model often failed to generate any text, produced repetitive content, or demonstrated a lack of planning ability by generating all text steps at once without interleaving images. Given these issues, reporting its scores would compromise the integrity of our evaluation.

CI-VID Dataset. For video narrative generation, we utilized the CI-VID dataset [33], which contains video clips with corresponding captions and inter-clip transition descriptions. To construct our training samples and mitigate potential black screen issues, we consistently selected the fifth frame from each video segment. The textual input for the initial frame is its corresponding clip's caption, while the guidance for all subsequent frames comes from the transition descriptions between clips. The first frame serves as the condition for generating the rest of the sequence. All training data from this dataset used a 480p anchor resolution (480×854), with the original video aspect ratio preserved.

8.3. Extended Application Scenarios (Q3)

To demonstrate the practical utility of our method, we apply Narrative Weaver to the domain of e-commerce advertising. By leveraging its dual capabilities in autonomous narrative planning and controllable consistency generation, our objective is to produce sequences of visual content that can serve as keyframes for complete advertising videos.

Evaluation Setup. For this application, we employ the same evaluation metrics as those used for Q1 in Sec. 5. We constructed a dedicated test set by randomly sampling 200 sequences from our generated e-commerce data.

Qualitative Comparison. While the main paper presented only a limited number of examples due to space constraints, this appendix provides a more comprehensive qualitative comparison. Fig. 15 showcases a side-by-side comparison between the results generated by Narrative Weaver and those from a leading image editing model. Since Qwen-Image-Edit lacks autonomous text generation capabilities, we supplied it with the same prompts generated by Narrative Weaver to ensure a fair comparison.

9. Experimental Details

Our multi-stage training strategy is designed to decouple Narrative Planning (Stage 1) from Visual Generation (Stages 2 and 3). This separation is enabled by a carefully designed attention mask that effectively freezes the narrative planning capability of the language model after Stage 1. Consequently, the subsequent stages can focus exclusively on enhancing coherent visual content generation without compromising the already-learned textual planning abilities.

We illustrate our training recipe using the e-commerce dataset (Q3) as a representative example. The detailed hyperparameters for each stage are provided in Tab. 5.

Table 5. Training recipe of Narrative Weaver.

Hyperparameters	Stage-1	Stage-2.1	Stage-2.2	Stage-3
Learning rate	1×10^{-5}	5×10^{-5}	1×10^{-5}	1×10^{-6}
LR scheduler	Constant	Cosine	Constant	Constant
Weight decay	0.0	0.0	0.0	0.0
Gradient norm clip	1.0	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.99, \epsilon = 1 \times 10^{-8}$)			
Loss type	CE	MSE	MSE	MSE
Warm-up steps	0	0	100	100
Training steps	32K	400K	48K	32K
Batch size	8	128	8	8
Module	Qwen2.5VL-3B	Learnable Query		Flux.1-Dev

Stage 1: Narrative Planning. In this stage, we train the Large Vision-Language Model (Qwen2.5VL-3B) on the task-specific dataset to master narrative and textual planning.

Stage 2: Bridging Language and Vision. This stage connects the planner with the visual generator and is divided into two sub-stages: Stage 2.1 (Connector Pre-training): We first pre-train the Learnable Queries on a large-scale public image-text dataset. The objective is to align these queries with the text embedding space of the visual generation model (Flux.1-Dev). Crucially, this pre-training is a one-time process. The resulting Learnable Queries can

be seamlessly reused as a plug-and-play module for various downstream fine-tuning tasks, eliminating the need for repeated training. Stage 2.2 (Task-specific Fine-tuning): The pre-trained Learnable Queries are then fine-tuned on the small, task-specific e-commerce dataset to adapt them to the specific domain.

Stage 3: Visual Generation Fine-tuning. Finally, we fine-tune the visual generation model (Flux.1-Dev) itself, further adapting it to the domain while keeping the other modules frozen.

This modular design makes the overall training process highly efficient for adapting to new tasks, as the Learnable Query pre-trained with large-scale data in Stage 2.1 can be seamlessly reused across diverse tasks without the need for repeated training. For all experiments presented in the main paper, we adopted a consistent image processing protocol. We used an anchor resolution of 480p (480×854), resizing all frames while preserving their original aspect ratio.

10. Additional Experimental Results

In this section, we provide extensive qualitative results to further substantiate the findings presented in the main paper. We offer more visualizations for each of our core experimental setups: controllable consistent generation (Q1), autonomous narrative generation (Q2), and the e-commerce application (Q3).

For controllable consistent generation (Q1), we first present an expanded set of qualitative results from Narrative Weaver in Fig. 12. These diverse examples further demonstrate the model's proficiency in maintaining high cross-frame consistency in subject identity, apparel, and background, while coherently evolving the narrative according to user prompts. Following this, Fig. 13 provides a side-by-side qualitative comparison with leading baseline models. This visualization highlights that while competitors can adhere to prompts, Narrative Weaver uniquely achieves a more cinematic and aesthetically pleasing quality in its outputs, showcasing superior handling of lighting, color, and composition.

For the task of autonomous narrative generation (Q2), additional examples are showcased in Fig. 14. These results underscore the model's robust planning capabilities, showing its ability to logically and creatively continue a story from a single initial prompt across a variety of scenarios.

Finally, regarding our e-commerce application (Q3), Fig. 15 presents a direct comparison against a leading editing model, Qwen-Image-Edit. This comparison illustrates our model's superior ability to preserve not only stylistic consistency with the reference image but also key semantic details required by the task, which is critical for real-world applications.

Showcase of a Full Advertising Production Pipeline. To demonstrate the practical utility of Narrative Weaver in a real-world production workflow, we produced complete, ready-for-deployment advertising videos, which are available in the supplementary materials. Our end-to-end pipeline is as follows:

First, we leverage Narrative Weaver to generate the core visual content: a sequence of high-quality, consistent keyframes that define the narrative. Next, we employ a Large Language Model (LLM) to create coherent and contextually appropriate shot descriptions or transition narratives for these keyframes. These image-text pairs are then fed into the Wan2.2 model for video synthesis, which generates a short video clip for each keyframe. Finally, all resulting video clips are concatenated to form a seamless, complete advertising video.

11. Detailed Ablation Results

To isolate and verify the contribution of Stage 3 in our training pipeline, we present a direct comparison between the full Narrative Weaver model and a variant trained without this final stage. As shown in Fig. 16, the model without Stage 3 can follow the core semantic instructions but fails to maintain strict visual consistency, leading to variations in subject appearance and details across frames.

The inclusion of Stage 3 rectifies this issue by enabling fine-grained control over the visual generation process. This results in a dramatic enhancement in the model's ability to preserve inter-frame consistency, which is critical for creating coherent visual narratives. This ablation clearly demonstrates that Stage 3 is essential for achieving the high-fidelity consistency that is a core strength of our method.

Figure 12. **Additional qualitative results of multi-frame narrative generation by Narrative Weaver.** Each row showcases a complete generation sequence. The leftmost column presents the user's initial input (a reference image and its description). The subsequent columns display the multi-frame visual narrative autonomously generated by our model, including both the synthesized images and their corresponding textual descriptions. These diverse examples highlight Narrative Weaver's proficiency in maintaining high cross-frame consistency—preserving subject identity, apparel, and key background elements—while coherently evolving the narrative through subtle changes in pose, expression, and camera perspective.

Figure 13. **Comparison with leading image editing models on multi-frame narrative generation.** While all models exhibit strong adherence to the text prompts and maintain high subject consistency, Narrative Weaver uniquely generates outputs with a more **cinematic and aesthetically pleasing quality**. Note our model's superior handling of lighting, color, and composition, which contributes to a more authentic, film-like visual narrative compared to the often more literal or digitally rendered feel of the baseline results.

The video depicts a cooking preparation process with half a kilogram of raw chicken. A person is handling the chicken pieces using a strainer over a large metallic pot placed on a wooden table, focusing on the initial steps of cooking.

The second clip continues with the chicken preparation. The person uses a strainer to transfer chicken pieces from a metal bowl to a large pot, actively participating in the cooking process.

In this clip, the chicken in the pot is being seasoned with a yellow liquid, likely a spice blend, and stirred with a spatula, focusing on the seasoning process.

Two people are mountain biking on a narrow, rocky dirt path in a dense forest. One wears a red helmet and camouflage shirt with a chest camera, riding a black bike. The other wears a white helmet and grey shirt on a dark bike, slightly ahead. They ride side by side, showcasing a mountain biking adventure.

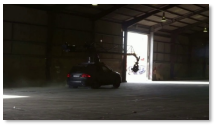
In the second clip, the mountain bikers continue along the rocky, narrow path. The rider in the red helmet leads initially, but the one in the white helmet gains a slight lead. Both maintain a steady pace, showcasing their skills in navigating the rugged terrain.

The video shows a thrilling moment in a baseball game at a sunny stadium. A Los Angeles Dodgers player in a white jersey slides towards home plate, while a catcher in a dark gray uniform prepares to catch the ball. The scene captures the excitement and anticipation of the game.

The video shows a high-speed stock car race with cars tightly packed on the track. A collision occurs, causing skidding and smoke, highlighting the chaos and unpredictability of racing. The fast pace and dynamic environment, transitioning from daylight to twilight, emphasize the excitement of racing.

The second clip shifts focus to a player in a blue and white uniform, number 13, running towards the outfield wall to catch a ball. This contrasts with the previous clip's focus on a player sliding into home plate. The scene highlights different players and actions, emphasizing the ongoing tension and urgency of the baseball game.

The second clip shifts focus to car number 97 in a close-up view, now stationary. The driver, wearing a helmet and racing suit, is visible inside. The car features a blue and yellow design with sponsor logos. The scene transitions from chaotic racing to a moment of preparation or contemplation, highlighting the driver's perspective and the car's details.

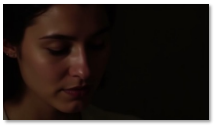


A black car with a crane and camera rig drives inside a dimly lit warehouse, kicking up dust on the concrete floor. The camera mount is extended for filming. As the car passes a large open doorway, natural light reveals industrial structures outside.

The video shows a close-up of a person with medium-length hair partially covering one ear. The focus is on their expressive hands near their face, highlighted against a dark background. The lighting emphasizes the hair and hands, creating a sharp contrast.

The black car continues as the focus in the second clip, now driving at a moderate speed on a paved road, kicking up dust. The camera rig and red taillights are still visible. The scene shifts from a dim warehouse to an open outdoor area with distant industrial buildings and a clear sky.

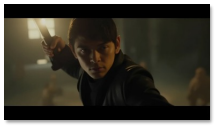
The second clip shows the same person, now focusing on their upper torso and shoulders. Their medium-length hair is visible, and their head is turned to the side. The lighting highlights their body contours and hair texture. The focus shifts from hand movements to the person's posture and hair.



The video shows a person standing confidently in a dimly lit area, holding two sheathed swords behind them. Dressed in a black outfit, their hair is tied back, emphasizing their determined posture. The setting suggests a training area or combat arena.

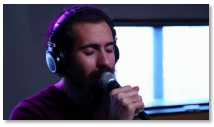
A man in a dark suit and fedora walks from a cozy living room with a fireplace and chessboard to a dining area. Another person is seated on the couch, and two children are at the dining table. The setting suggests a residential home during the evening.

The second clip shows the an individual now in action, wearing a black leather jacket and wielding a sword. The character demonstrates agility and skill in a fast-paced sequence through a corridor, leaping and swinging the sword.



The man in the dark suit remains the focus, now in a room with a wooden cabinet and a large painting. He turns his head and shifts his gaze, contemplating the artwork. The setting changes from a living room and dining area to this new room.

A person in a slightly dirty white uniform is engaged in physical work. Their left arm is outstretched over a flat surface, while their right hand holds a tool to spread a substance. The uniform is stained, indicating the nature of the task.



A person in large Audio-Technica headphones is singing into a handheld microphone, wearing a dark maroon long-sleeve shirt indoors. They hold the microphone firmly with their right hand and gesture expressively with their left. Subtle head movements show engagement in the performance.

The person is now working with a piece of meat, carefully seasoning or marinating it. Their dirty hands apply and adjust seasonings on the meat laid on a flat surface. The focus shifts from a general task to meat preparation, with hand application replacing the tool.

The second clip shows the person continuing their vocal performance with over-ear headphones and a microphone. The microphone has a visible cable, and the person's hand is holding the microphone stand. The focus is closer on the microphone and hand, with less visibility of their upper body.

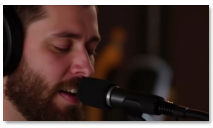


Figure 14. **Qualitative examples of autonomous story continuation by Narrative Weaver.** Given only the first frame and text as input for each sequence, Narrative Weaver autonomously plans and generates a coherent multi-frame continuation. The diverse examples—from procedural tasks like cooking to dynamic events like sports—showcase the model's robust planning capabilities. Its ability to logically advance a narrative by continuing actions, introducing new elements, or shifting focus highlights its understanding of storytelling beyond simple image editing.

Figure 15. **Comparison with a leading editing model, Qwen-Image-Edit.** Narrative Weaver not only demonstrates superior stylistic consistency with the conditional image but also excels in preserving key semantic details required by the task. In contrast, Qwen-Image-Edit exhibits noticeable failure modes: it struggles with inconsistent color tones between frames (upper example) and introduces a warm color cast that deviates from the style of the reference image (lower example).

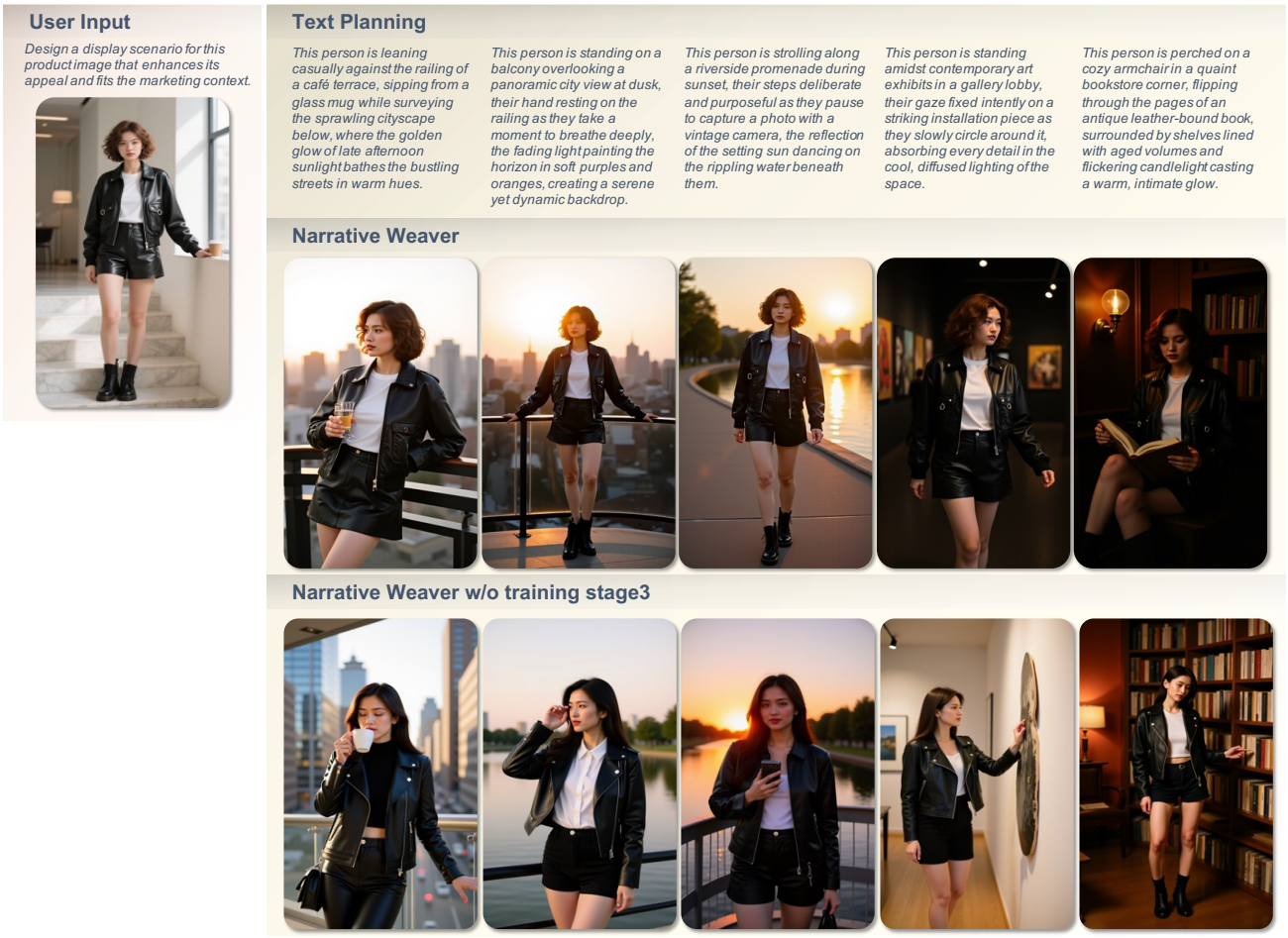


Figure 16. **Ablation Study Results.** Comparison between Narrative Weaver and its variant without Stage 3 training demonstrates that: (1) the first two stages establish fundamental semantic alignment; (2) Stage 3 significantly enhances inter-frame consistency; and (3) Stage 3 is crucial for imparting fine-grained control, as the variant without it produces images that deviate significantly from the reference.

Table 6. **Computational cost (TFLOPs) as a function of the number of generated keyframes.** Our approach demonstrates significantly better scalability compared to a vanilla self-attention baseline. The computational cost of our method grows linearly with the sequence length, unlike the vanilla implementation, whose cost grows quadratically, making our method far more efficient for longer sequences.

Implementation	Keyframe Num	1	2	3	4	5	6	7	8	9	10	11	12	16	20
Vanilla	TFLOPs	82	230	450	744	1112	1553	2068	2656	3318	4053	4862	5744	10008	15449
Ours		82	165	248	331	441	497	580	663	746	829	912	995	1077	1160

12. Additional Efficiency Analysis

The superior efficiency of Narrative Weaver, as demonstrated in Tab. 6, stems from a fundamental architectural choice. Specifically, our approach delegates the task of establishing cross-frame coherence to the Multimodal Large Language Model (MLLM) planning stage. As a result, the input token sequence for the Diffusion Transformer (DiT) remains constant in length, regardless of the number of keyframes being generated.

In contrast, vanilla implementations [27, 28] must maintain coherence within the DiT itself. This is typically achieved by concatenating the latent representations of all preceding frames and processing them simultaneously, causing the sequence length to grow with each new frame. This architectural difference directly leads to the quadratic explosion in computational complexity observed for the vanilla method in Tab. 6, whereas our approach maintains a highly efficient, near-linear scaling.

13. Limitations and Future Works

In this paper, we introduced Narrative Weaver, a unified framework capable of fine-grained control, long-range consistency preservation, and autonomous narrative planning. While the architecture is theoretically capable of generating any form of visual content, our current work presents a preliminary implementation focused exclusively on images.

We acknowledge that focusing on long-range *image* consistency is a pragmatic choice, largely dictated by current resource constraints. A critical and promising direction for future work is to extend Narrative Weaver to ensure consistency across multiple *video clips*. This extension is vital because video introduces crucial elements of temporal consistency, including coherent character motion and logical camera movements (cinematography), which are not captured in static images. We leave this ambitious extension for future investigation.

Furthermore, our research highlights a significant challenge in the field: the scarcity of high-quality datasets designed for controllable, long-range consistent content generation. To address this gap, we constructed a new dataset tailored to the e-commerce domain. However, the development of similar large-scale, diverse datasets for broader application domains remains a critical need for advancing research in this area and represents another important avenue for future work.