# BrandFusion: A Multi-Agent Framework for Seamless Brand Integration in Text-to-Video Generation

**Zihao Zhu**[1] **Ruotong Wang**[1] **Siwei Lyu**[3] **Min Zhang**[4] **Baoyuan Wu**[1,2*]
[1]The Chinese University of Hong Kong, Shenzhen [2]Shenzhen Loop Area Institute
[3]State University of New York at Buffalo [4] Harbin Institute of Technology
{zihaozhu, ruotongwang1}@link.cuhk.edu.cn;
siweilyu@buffalo.edu; zhangmin2021@hit.edu.cn; wubaoyuan@cuhk.edu.cn

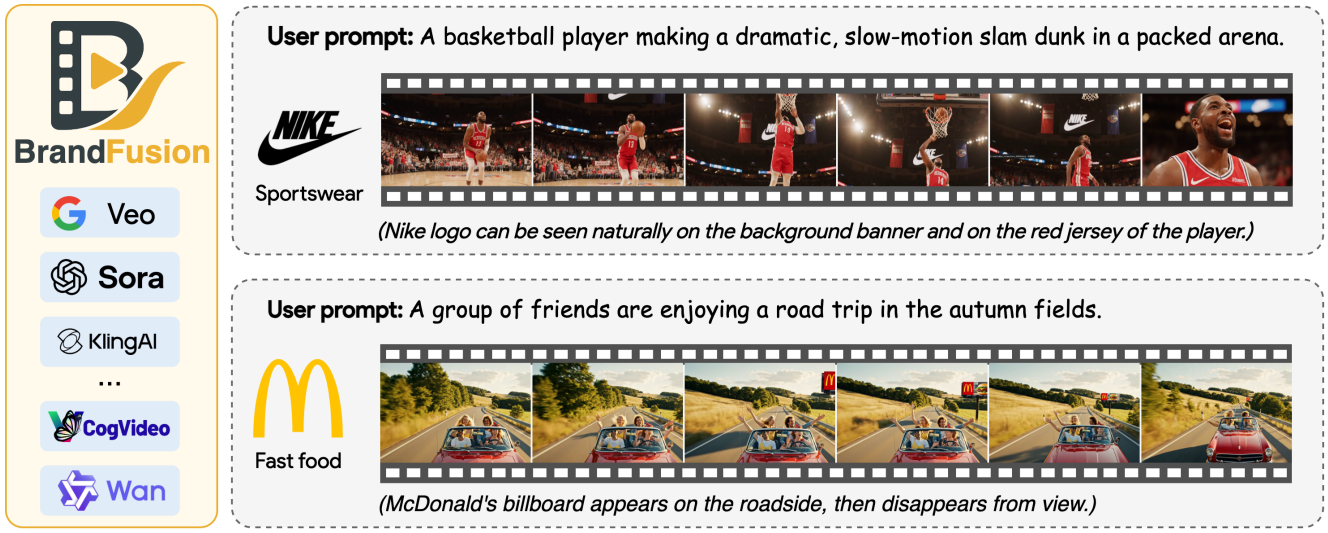


Figure 1. **Examples of seamless brand integration by BrandFusion.** Given user prompts like basketball games and cyberpunk street scenes, our framework naturally incorporates brands (Nike on jersey/banner, Coca-Cola on billboard) into generated videos. BrandFusion can integrate with multiple T2V models and handle diverse brands while preserving user intent and ensuring natural visual coherence.

## Abstract

*The rapid advancement of text-to-video (T2V) models has revolutionized content creation, yet their commercial potential remains largely untapped. We introduce, for the first time, the task of seamless brand integration in T2V: automatically embedding advertiser brands into prompt-generated videos while preserving semantic fidelity to user intent. This task confronts three core challenges: maintaining prompt fidelity, ensuring brand recognizability, and achieving contextually natural integration. To address them, we propose **BrandFusion**, a novel multi-agent framework comprising two synergistic phases. In the **offline phase** (advertiser-facing), we construct a Brand Knowledge Base by probing model priors and adapting to novel brands via lightweight fine-tuning. In the **online phase** (user-facing), five agents jointly refine user prompts through iterative refinement, leveraging the shared knowledge base and real-time contextual tracking to ensure brand visibility and semantic alignment. Experiments on 18 established and 2 custom brands across multiple state-of-the-art T2V models demonstrate that BrandFusion significantly outperforms baselines in semantic preservation, brand recognizability, and integration naturalness. Human evaluations further confirm higher user satisfaction, establishing a practical pathway for sustainable T2V monetization.*

## 1. Introduction

The rapid advancement of text-to-video (T2V) generation models, such as Veo [8], Sora [28], and Kling [21], has

[1]Corresponding Author.

revolutionized content creation, enabling users to synthesize high-fidelity videos from natural language descriptions [1, 2, 11, 20, 24, 33, 36, 39, 47]. However, as these technologies transition to commercial products, establishing sustainable monetization models remains an open challenge given the substantial computational costs involved. In this work, we introduce, for the first time, the task of **seamless brand integration in T2V generation**: automatically embedding advertiser brands into prompt-generated videos while preserving semantic alignment with user intent.

Unlike intrusive traditional advertising that disrupts user experience [4, 6, 26, 29–31], our proposed task aims to naturally embed brand elements within generated content, preserving the user's creative intent and aesthetic control while achieving brand visibility. This paradigm opens a promising pathway for sustainable T2V services: advertisers gain organic exposure through contextually relevant placements, service providers establish viable revenue streams, and users continue to receive high-quality content without explicit interruption. The task's success hinges on a delicate balance: brands must be recognizable yet unobtrusive, visually prominent yet contextually harmonious, and semantically coherent with the user's original prompt.

However, seamlessly integrating brands into T2V generation presents formidable challenges that extend far beyond simple text manipulation. First, maintaining *semantic alignment* [48, 49] requires that brand integration preserves the user's original creative intent, including key subjects, actions, and stylistic preferences, as any deviation risks user dissatisfaction and service abandonment. Second, ensuring *brand visibility* demands that brand elements appear recognizably and identifiably within generated videos, as insufficient prominence fails to deliver advertising value to brand owners. Third, achieving *natural integration* necessitates that brands appear organically within scene contexts, avoiding jarring or out-of-place elements that disrupt visual coherence. These three objectives often conflict: overly prominent placement may compromise naturalness, while prioritizing subtlety may sacrifice visibility. Furthermore, the diversity of user prompts spanning countless scenarios, combined with the variety of brands ranging from well-established entities to emerging startups across categories such as beverages, automobiles, and apparel, creates a vast combinatorial space. Rule-based approaches inevitably fail to generalize, frequently producing abrupt or incoherent outputs that satisfy neither users nor advertisers.

To address these challenges, we propose **BrandFusion**, a multi-agent framework for seamless brand integration in T2V generation. Our key insight is that effective integration requires both comprehensive brand knowledge and sophisticated contextual reasoning, capabilities that emerge naturally from multi-agent collaboration. As shown in Figure 3, it operates through two synergistic phases: **(1)** An *offline phase* (advertiser-facing) that constructs a Brand Knowledge Base by probing T2V models' prior knowledge of brands and selectively performing lightweight fine-tuning for novel brands, which serves as the foundation for subsequent online integration. **(2)** An *online phase* (user-facing) that employs five specialized agents, namely brand selector, strategy generator, prompt refiner, critic, and experience learner, to collaboratively refine prompts through iterative refinement. In this multi-agent system, agents coordinate by leveraging the shared Brand Knowledge Base (storing brand profiles, adapters, and accumulated experiences) and a real-time Session Context that tracks the current generation state for adaptive prompt adjustment. As illustrated in Figure 1, BrandFusion generates videos with natural brand presence across diverse scenarios. Extensive experiments on 18 well-known brands and 2 custom brands across multiple T2V models demonstrate that our framework consistently outperforms baselines in semantic alignment, brand visibility, and integration naturalness, with human evaluation confirming superior user satisfaction.

Our main contributions are three-fold: **(1)** We introduce seamless brand integration in T2V generation as a novel task, accompanied by comprehensive evaluation protocols. **(2)** We propose BrandFusion, a multi-agent framework featuring systematic offline brand knowledge construction and online collaborative prompt refinement. **(3)** We conduct extensive experiments on both established and novel brands across multiple T2V models, achieving state-of-the-art performance with human evaluations confirming superior user satisfaction.

## 2. Related Work

**Prompt Optimization for Video Generation.** The quality of generated videos critically depends on the effectiveness of input prompts, and prompt optimization aims to enhance user-provided rough prompts into detailed, model-preferred descriptions. Training-based methods [3, 9, 19, 37, 41] employ supervised fine-tuning and reinforcement learning to refine prompts through reward-guided optimization, assessing metrics such as image-text alignment and compositional fidelity. Multi-agent frameworks [13, 15, 25, 38, 42, 43, 45, 46] decompose prompt refinement into specialized agents that collaboratively handle tasks like scene enrichment, verification, and correction through iterative workflows. Unlike existing methods that solely focus on generation quality, our work also seamlessly integrates brand elements in the video, thereby establishing monetization pathways for T2V service providers.

**Brand Integration in Generative Models.** In text-to-image diffusion models, model customization techniques such as DreamBoot [32] and Textual Inversion [5] enable models to synthesize novel entities through lightweight fine-tuning. Recent work has attempted brand embed-

ding through adversarial approaches: Silent Branding Attack [18] employs data poisoning to covertly inject brand logos into training data, enabling models to generate branded content without explicit prompts, while BAGM [34] implements backdoor attacks across different model components to manipulate outputs with brand elements. These methods operate covertly without user awareness, prioritizing stealth and concealment. Different from these approaches, we are the first to introduce seamless brand integration for text-to-video generation, naturally incorporates brands into user-requested videos while preserving semantic fidelity.

## 3. Seamless Brand Integration in Text-to-Video Generation

### 3.1. Preliminaries

Text-to-video (T2V) generation models synthesize video sequences from natural language descriptions. Given a text prompt $\mathcal{P}$, a T2V model generates a corresponding video $\mathcal{V}$. Contemporary T2V models leverage diffusion-based architectures where text prompts are encoded into semantic embeddings via pretrained encoders (*e.g.*, CLIP, T5) that guide video synthesis through cross-attention mechanisms. The quality and specificity of the prompt critically influence the generated video's semantic alignment with user intent.

### 3.2. Task Description

The rapid advancement of T2V models has opened unprecedented opportunities for automated content creation, particularly in advertising production [23, 26]. We introduce the task of **seamless brand integration for T2V generation**, a novel problem that naturally incorporates brand elements into user-requested videos while preserving semantic fidelity and user intent.

**Task Definition.** Given a user-provided text prompt $\mathcal{P}_u$ and a advertiser-provided brand profile $\mathcal{B}$, the task's goal is to generate an optimized prompt $\mathcal{P}_{\text{opt}}$ that guides the T2V model to produce a video $\mathcal{V}$ integrated with the brand while satisfying three critical constraints: *(1) Semantic Fidelity*: The generated video must faithfully reflect the user's original intent, preserving key semantic elements and narrative flow. *(2) Brand Presence*: The brand elements must be recognizably integrated into the video with clear visibility and identifiability. *(3) Natural Integration*: The brand should appear organically within the scene context, avoiding obtrusiveness or semantic incongruity.

### 3.3. Application Scenarios

This task demonstrates practical real-world application prospects in commercial ecosystems connecting brand owners, T2V service providers, and end users. As illustrated in Figure 2, the ecosystem operates through the following workflow: brand owners first pay T2V service providers to register their brand profiles and enable advertising; the T2V providers then seamlessly integrate brand information when end users submit creative requests; subsequently, users receive high-quality branded videos that balance creative intent with brand visibility; ultimately, brands achieve organic exposure, establishing sustainable monetization channels for service providers.

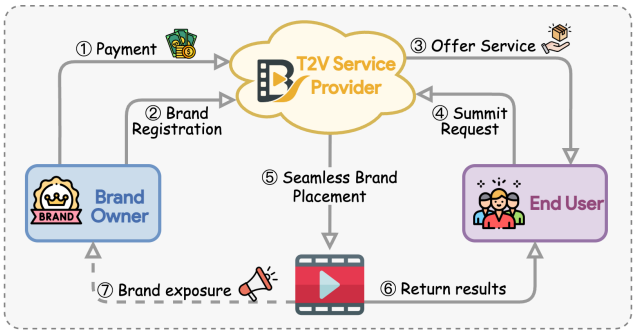


Figure 2. Ecosystem of brand integration in T2V generation.

## 4. Methodology

### 4.1. Overview of BrandFusion

In this paper, we propose **BrandFusion**, a multi-agent system for seamless brand integration in T2V generation. As illustrated in Figure 3, it consists of two synergistic phases: (1) *Offline Brand Knowledge Base Construction*, which adaptively builds brand knowledge through prior knowledge probing and selective model adaptation, and (2) *Online Multi-Agent Brand Integration*, which employs five specialized agents to intelligently optimize brands while preserving semantic fidelity. A centralized Brand Knowledge Base serves as long-term memory, accumulating knowledge and successful experiences across generations.

### 4.2. Phase I: Offline Brand Knowledge Base Construction

As shown in Figure 3 (a), the offline phase systematically prepares brand knowledge for efficient online integration. Given a brand profile $\mathcal{B} = \{\mathcal{N}, \mathcal{C}, \mathcal{R}, \mathcal{D}\}$ from advertisers comprising brand name, category, reference images, and description, we first probe the T2V model's prior knowledge of the brand. Brands with sufficient prior knowledge are directly registered in the Brand Knowledge Base, while those lacking knowledge undergo model-level adaptation to generate brand-specific adapters. The Brand Knowledge Base maintains comprehensive information for each brand, including knowledge type, adapter weights (if applicable), reference visual patterns, and an experience pool of successful integration cases, serving as the foundation for subsequent online integration.

**Prior Knowledge Probing.** To determine whether the T2V

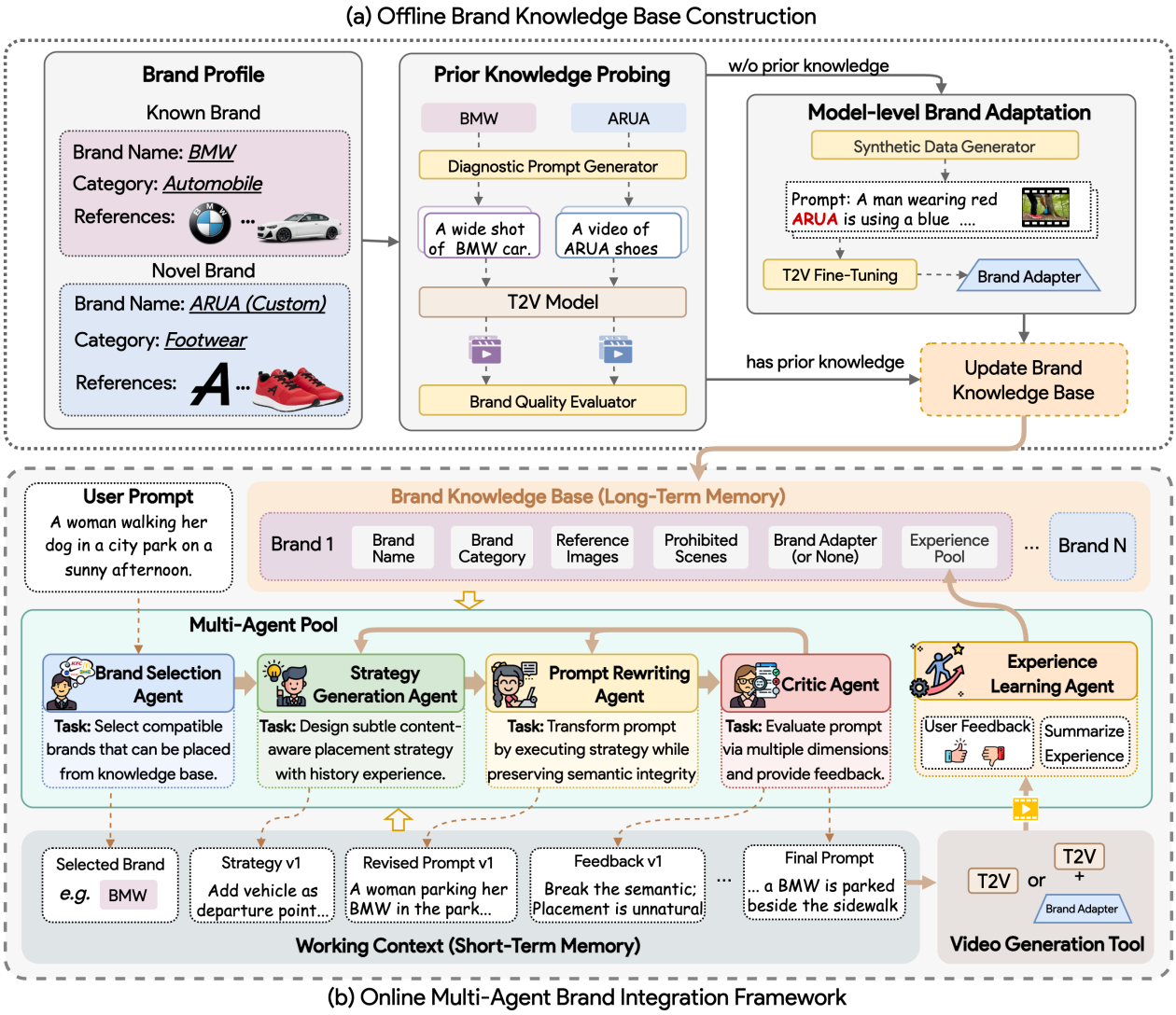


Figure 3. Overview of the BrandFusion framework: (a) Offline phase builds brand knowledge through probing and adaptation. (b) Online phase employs five collaborative agents for semantic-preserving brand integration with continuous learning.

model possesses adequate knowledge of a given brand, we employ a systematic probing procedure. A *diagnostic prompt generator* first creates diverse test prompts that explicitly mention the brand across various contexts (*e.g.*, *"A wide shot of a BMW car"*). The model then generates videos from these prompts. A *brand quality evaluator* examines each video to detect whether brand elements are accurately present and visually correct. If more than 70% of the videos successfully generate the brand with recognizable features, the brand is marked as having sufficient prior knowledge and directly stored in the knowledge base.

**Model-level Brand Adaptation.** For brands lacking prior knowledge, we perform lightweight model adaptation to inject specific brand knowledge. A *synthetic data generator* constructs training data by: (1) generating diverse prompts mentioning the brand with a special trigger token (*e.g.* the brand name); (2) synthesizing corresponding videos using reference images $\mathcal{R}$ and text-to-image models to create initial frames; (3) extending to full videos through image-to-video techniques. This process yields a synthetic dataset $\mathcal{D}_{\text{syn}} = \{(\mathcal{P}_j, \mathcal{V}_j)\}_{j=1}^{M}$. We fine-tune the T2V model using LoRA on $\mathcal{D}_{\text{syn}}$, producing a brand-specific adapter $\mathcal{A}_{\mathcal{B}}$. The adapted model can then accurately generate videos containing the brand when the trigger token appears in prompts. Both the adapter weights and metadata are stored in the Brand Knowledge Base, enabling efficient retrieval during online integration.

### 4.3. Phase II: Online Multi-Agent Brand Integration Framework

Given a user prompt $\mathcal{P}_u$, the online phase employs a multi-agent system to seamlessly integrate brands while preserving semantic integrity and user intent.

**Multi-Agent System Structure.** Our framework comprises five specialized agents, each agent is powered by large language models and performs distinct functions:

- **Brand Selection Agent**: The first agent queries the Brand Knowledge Base to select the most compatible brand for $\mathcal{P}_u$ by considering semantic compatibility between the prompt's scene/context and the brand's typical application scenarios; The agent outputs the selected brand $\mathcal{B}^*$ and its profile including whether it requires adapter.
- **Strategy Generation Agent**: This agent aims to design subtle and context-aware integration strategies that balance semantic preservation with brand visibility. It analyzes the user prompt's scene characteristics to determine optimal integration strategy $\mathcal{S}$. In addition, it can query the experience pool to analysis historical successful strategies from similar scenarios, avoiding unsuccessful approaches from past failures.
- **Prompt Rewriting Agent**: This agent transforms $\mathcal{P}_u$ into an optimized prompt $\mathcal{P}'$ by executing the strategy while adhering to four core principles: *Semantic Preservation* maintaining original subject and actions, *Natural Integration* incorporating brand elements as natural scene components, *Logical Consistency* preserving temporal and causal coherence, and *Style Consistency* following T2V prompt conventions.
- **Critic Agent**: This agent performs multi-dimensional evaluation of $\mathcal{P}'$, assessing semantic fidelity to $\mathcal{P}_u$, brand clarity and recognizability, integration naturalness and fluency, and expected generation effectiveness. Based on the evaluation, it decides to accept the prompt, revise the prompt or replan the strategy.
- **Experience Learning Agent**: Once the prompt is accepted, the system generates video with the T2V model, conditionally loading the brand adapter if available. This agent then collects user feedback and abstracts it into experiences: positive feedback yields successful patterns, while negative feedback records failed ones. Both are stored in the experience pool, enabling continuous system improvement through closed-loop learning.

**Inter-Agent Collaboration Mechanism.** The agents collaborate through dual memory: the *brand knowledge base* (long-term) stores brand profiles, adapters, and historical experiences, while the *working context* (short-term) tracks current session state. Brand selection retrieves $\mathcal{B}^*$ from long-term memory; Strategy generation leverages historical experiences to formulate $\mathcal{S}$; Prompt rewriting produces $\mathcal{P}'$; Critic evaluates and provides feedback, triggering iterative refinement until acceptance. All intermediate results are logged in working context. Finally, Experience learning abstracts the completed integration and writes it back to long-term memory, enabling continuous improvement through closed-loop learning.

## 5. Experiments

### 5.1. Experimental Setup

**Brand Integration Benchmark.** To comprehensively evaluate BrandFusion's integration capabilities, we construct a two-tiered benchmark. For **known brands** with existing model priors, we curate 18 well-established brands spanning 7 industry categories. For each brand, we manually construct 15 diverse prompts with different propmt-brand compatibility: *High Match* prompts where the brand naturally fits, *Medium Match* prompts where the brand can reasonably appear, and *Low Match* prompts representing challenging scenarios with low relevance. This yields 270 (brand, prompt) pairs across varying integration difficulty levels. For **novel brands** lacking prior knowledge, we design two fictional brands: a sportswear brand "ARUA" and an beverage brand "FreshWave". Each brand is equipped with logo and product images and corresponding test prompts.

**T2V Model Selection.** For known brands, we evaluate the multi-agent framework on three commercial T2V models: Veo3 [8], Sora2 [28], and Kling2.1 [21]. For novel brands, we fine-tune adapters on three open-source models: Wan2.1-1.3B, Wan2.2-5B [35], and CogVideoX-5B [44].

**Baeline Methods.** We compare BrandFusion against three baseline approaches: (1) *Direct Append* naively appends the brand name to the prompt end; (2) *Template-based Rewriting* uses fixed templates to insert brands into prompts; (3) *Single Rewriting* employs LLM for one-pass prompt rewriting.

**Evaluation Metrics.** To comprehensively assess brand integration quality, we establish a multi-dimensional evaluation framework encompassing automated metrics and human assessments across three critical aspects: video generation quality, semantic fidelity to user intent, and brand integration quality.

- **Video Quality.** We adopt the VBench Quality Score (VBench-Quality) [16, 17, 50], a widely-used comprehensive metric aggregating multiple dimensions from temporal quality and frame-wise quality perspectives. Higher score indicates better overall video generation quality.
- **Semantic Fidelity.** To measure how well generated videos preserve the user's original intent, we employ three complementary metrics: (1) *VQAScore* leverages visual question answering to assess semantic preservation by generating key questions from the original prompt and evaluating whether the generated video provides cor-

Table 1. Main results on well-known brands across three T2V models. BrandFusion achieves comparable visual quality while significantly outperforming baselines on semantic fidelity and brand integration quality. Higher scores indicate better performance.

| T2V | Method | Video Quality | Semantic Fidelity | | | Brand Integration Quality | |
|---|---|---|---|---|---|---|---|
| | | VBench-Quality | CLIPScore | VQAScore | LLMScore | Brand Presence Rate (BPR) | Naturalness Score (NS) |
| Veo3 | Direct Append | 0.8112 | 0.2671 | 0.8342 | 0.7821 | 0.7221 | 2.8300 |
| | Template Rewriting | 0.8267 | 0.2842 | 0.8756 | 0.9234 | 0.8845 | 3.1200 |
| | Single Rewriting | 0.8289 | 0.2956 | 0.8891 | 0.9412 | 0.8968 | 3.9000 |
| | BrandFusion (Ours) | 0.8283 | 0.3274 | 0.9098 | 0.9556 | 0.9474 | 4.7000 |
| Sora2 | Direct Append | 0.7945 | 0.2645 | 0.8298 | 0.8756 | 0.6434 | 2.7100 |
| | Template Rewriting | 0.8033 | 0.2868 | 0.8712 | 0.9187 | 0.7845 | 3.8400 |
| | Single Rewriting | 0.8029 | 0.2968 | 0.8867 | 0.9368 | 0.8278 | 3.9800 |
| | BrandFusion (Ours) | 0.8031 | 0.3177 | 0.9231 | 0.9875 | 0.9066 | 4.6000 |
| Kling2.1 | Direct Append | 0.7754 | 0.2634 | 0.8276 | 0.8742 | 0.6989 | 2.6500 |
| | Template Rewriting | 0.7803 | 0.2855 | 0.8689 | 0.9165 | 0.7812 | 3.7300 |
| | Single Rewriting | 0.7883 | 0.2951 | 0.8823 | 0.9351 | 0.8145 | 3.8400 |
| | BrandFusion (Ours) | 0.7818 | 0.3165 | 0.9208 | 0.9853 | 0.8834 | 4.4800 |

rect answers [14, 22, 51]; (2) *CLIPScore* measures alignment between generated video and original prompt using CLIP embeddings [10, 40], computing average similarity across frames; (3) *LLMScore* employs multimodal large language models to assess semantic consistency. Higher scores indicate better semantic preservation.

- **Brand Integration Quality.** We evaluate from two complementary perspectives: (1) *Brand Presence Rate (BPR)* is a binary metric indicating whether brand elements are successfully detected in the generated video using multimodal LLMs; (2) *Naturalness Score* assesses integration quality through three LLM-evaluated factors on a 1-5 scale, including contextual fit , visual blend, and non-intrusivenesst. The final naturalness score is averaged across these three dimensions. Higher values indicate more natural integration.

**Implementation Details.** We implement all agents using GPT-5 [27] with temperature 0.7. For adapter training, we employ LoRA [12] with rank 32, learning rate $1 \times 10^{-4}$, and train for 10 epochs. For each custom brand, we synthesize 100 training videos by first generating diverse prompts with trigger tokens, then creating initial frames via Nano-Banana [7] conditioned on reference images, and finally extending to full videos using Veo-3. The prompts of each agent are provided in the Appendix.

## 5.2. Main Results

**Results on Known Brands.** We first evaluate BrandFusion on the known brands in our benchmark. To ensure fair comparison across varying difficulty match levels and baselines, we configure brand selection agent to output the predetermined brand for each prompt. Table 1 presents results averaged across all test cases. BrandFusion achieves VBench quality scores comparable to baselines, indicating that brand integration does not compromise generation quality. However, BrandFusion significantly outperforms all baselines on semantic fidelity and naturalness scores, demonstrating that our framework successfully balances semantic preservation with natural integration. Notably, BrandFusion maintains high brand presence while achieving superior naturalness, indicating effective brand visibility without sacrificing scene coherence. Single-LLM Rewriting lacks iterative refinement, resulting in lower naturalness. Template-based and Direct Append methods achieve high brand presence but often produce abrupt results.

Table 2. Results on novel brands with model-level adaptation.

| Brand | T2V Model | CLIPScore | LLMScore | BPR | NS |
|---|---|---|---|---|---|
| ARUA | Wan2.1-T2V-1.3B | 0.2934 | 0.9134 | 0.9312 | 3.94 |
| | Wan2.2-T2V-5B | 0.3056 | 0.9487 | 0.9556 | 4.18 |
| | CogVideoX1.5-5B | 0.2789 | 0.8923 | 0.8745 | 3.27 |
| FreshWave | Wan2.1-T2V-1.3B | 0.2898 | 0.9178 | 0.8267 | 3.98 |
| | Wan2.2-T2V-5B | 0.3012 | 0.9523 | 0.9312 | 4.39 |
| | CogVideoX1.5-5B | 0.2723 | 0.8867 | 0.7723 | 3.19 |

**Results on Novel Brands.** We evaluate the effectiveness of brand adaptation for two novel brands. As shown in Table 2, Wan2.2-T2V-5B consistently achieves the best performance with highest BPR and NS scores, attributed to its larger model capacity and stronger generation capabilities, while Wan2.1-T2V-1.3B maintains reasonable effectiveness despite its smaller size. Notably, all three models achieve high brand presence rates and naturalness scores, validating our adapter training's effectiveness. The consistent performance across diverse brand categories (footwear vs. beverage) confirms the generalizability of our synthetic data generation and LoRA fine-tuning approach, successfully enabling T2V models to generate novel brands with visual fidelity and natural scene integration even when en-

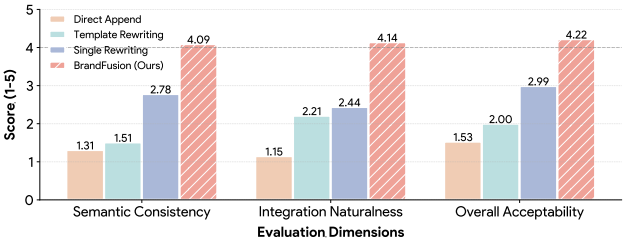


Figure 4. Human evaluation results on semantic fidelity, integration naturalness, and overall acceptability.

tirely absent from pre-training data.

**Human Evaluation.** To validate our approach from the user perspective, we conduct a user study with 10 participants evaluating videos generated by the Veo3 model[1]. Participants assess videos across three dimensions using a 1-5 Likert scale: (1) *Semantic Fidelity* — preservation of original prompt intent; (2) *Integration Naturalness* — how naturally the brand is incorporated; and (3) *Overall Acceptability* — satisfaction with the generated video. As shown in Figure 4 , BrandFusion consistently achieves the highest scores across all dimensions. This confirms that our framework successfully balances brand integration with user satisfaction, while baseline methods compromise user experience through semantic deviation or unnatural placements.

## 6. Analysis

### 6.1. Performance Across Different Match Level

To assess robustness across varying scene-brand compatibility, we analyze performance stratified by match level. Table 3 reveals that BrandFusion maintains consistently strong performance across all compatibility levels, achieving high semantic fidelity and naturalness even in challenging Low Match scenarios. In contrast, baseline methods exhibit sharp degradation: Template-based Rewriting drops from NS 4.01 in high match to 1.38 in low match, while Direct Append shows significant decline in low match cases. BrandFusion gracefully handles difficult scenarios, maintaining strong semantic preservation while finding creative, context-aware integration strategies. This advantage validates the effectiveness of iterative multi-agent collaboration over one-pass optimization or fixed heuristics.

### 6.2. Performance Across Prompt Scene Categories

To evaluate BrandFusion's adaptability across diverse real-world scenarios, we analyze performance on seven major scene categories aggregated from our 14 prompt scene types generated by Veo3. As illustrated in Figure 5, BrandFusion demonstrates consistent superiority over all baseline methods. Our method achieves particularly strong results in everyday contexts such as Urban Scenes, Social & Home Life, where natural brand placement opportunities are abundant. While Temporal Themes, such as Sci-Fi and Historical scenes, present greater challenges due to the need to reconcile modern brands with temporally distinct contexts, BrandFusion still outperforms all baselines. These results validate that our framework enables effective cross-category generalization, maintaining high-quality brand integration regardless of scene complexity.

[1]This study was approved by our institution's IRB.

Table 3. Performance across different prompt-brand match levels.

| Method | High Match | | Medium Match | | Low Match | |
|---|---|---|---|---|---|---|
| | LLMScore | NS | LLMScore | NS | LLMScore | NS |
| Direct Append | 0.8234 | 3.95 | 0.7876 | 3.21 | 0.7353 | 1.33 |
| Template Rewriting | 0.9456 | 4.01 | 0.9178 | 3.97 | 0.9068 | 1.38 |
| Single Rewriting | 0.9512 | 4.68 | 0.9489 | 4.05 | 0.9235 | 2.97 |
| BrandFusion (Ours) | 0.9734 | 4.90 | 0.9601 | 4.78 | 0.9333 | 4.42 |

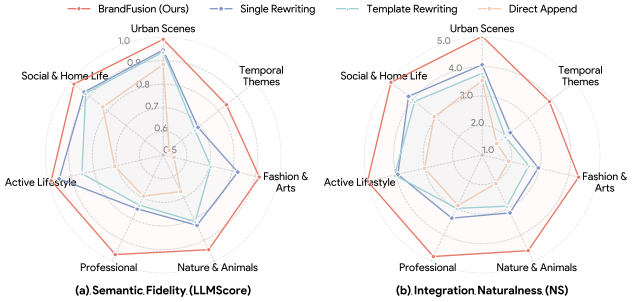


Figure 5. Comparison across different prompt scene categories.

### 6.3. Performance Across Brand Categories

To evaluate BrandFusion's generalizability across commercial domains, we analyze performance on seven brand categories spanning Food & Beverage, Technology, Transportation, Apparel, Beauty, Home, and Health & Wellness. As shown in Figure 7, BrandFusion consistently outperforms all baselines in both semantic fidelity and integration naturalness. Notably, everyday brands such as Apparel & Footwear achieve the highest integration quality, attributed to their natural association with human subjects and flexible placement across diverse scenarios. Baseline methods show considerable performance variation, with Direct Append particularly struggling in specialized domains. While categories like Technology and Transportation present greater integration challenges due to more constrained contextual requirements, BrandFusion maintains relative higher performance, validating the effectiveness of our adaptive multi-agent framework across diverse commercial domains.

### 6.4. Experience Learning Effectiveness

To validate the effectiveness of experience learning mechanism, we construct 100 diverse prompts for BMW brand integration, process them sequentially, and divide them into


Figure 6. Case study examples of brand integration for novel brands using Wan2.2-T2V-5B. ARUA sportswear (top) and FreshWave beverage (bottom) are seamlessly integrated into diverse scenarios, demonstrating natural brand placement while preserving user intent.

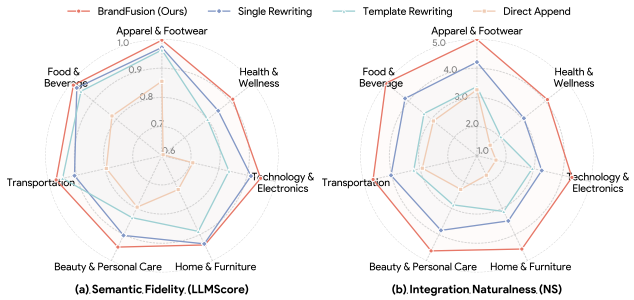


Figure 7. Comparison across seven brand categories.

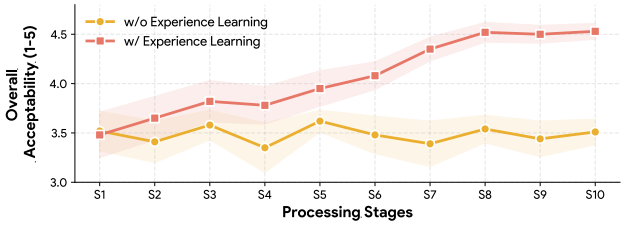


Figure 8. Experience learning effectiveness over sequential stages.

10 groups to compute average overall acceptability within each group. As illustrated in Figure 8, BrandFusion with experience learning shows a clear upward trend across stages, while the baseline without experience learning remains relatively flat, demonstrating that the Experience Learning Agent successfully accumulates knowledge from previous integrations and continuously refines strategies.

### 6.5. Efficiency Analysis

While BrandFusion employs a multi-agent framework with iterative refinement, the online optimization process maintains reasonable efficiency. Our analysis reveals that the framework requires an average of 7.4 LLM calls per prompt, with an average latency of 16 seconds. This overhead is acceptable compared to actual video generation: Wan2.2-T2V-5B requires at least 120 seconds to generate a single video on an A100 GPU, meaning prompt optimization accounts for about 11% of the pipeline. Combined with the substantial quality improvements demonstrated in our experiments, this makes BrandFusion viable for production deployment where video generation remains the dominant latency factor.

### 6.6. Case Studies

Figure 6 illustrates representative examples of brand integration for novel brands using Wan2.2-T2V-5B. For the ARUA sportswear brand, the red shoes appear naturally on the rider's feet in a lawn mowing scene, maintaining clear visibility while preserving the activity context. For FreshWave beverage, the green bottle integrates seamlessly on the table during a backyard barbecue gathering, appearing as an organic element among other items. Both examples demonstrate BrandFusion's capability to handle novel brands absent from pre-training data, achieving recognizable brand presence while maintaining semantic fidelity and visual naturalness. More examples are shown in Appendix.

## 7. Conclusion

In this paper, we introduce seamless brand integration for text-to-video generation, a novel task that embeds brands into videos while preserving semantic fidelity to the user's original intent. We propose BrandFusion, a multi-agent framework featuring offline brand knowledge construction and online collaborative prompt refinement through five specialized agents. Extensive experiments on 18 established and 2 novel brands across multiple state-of-the-art T2V models demonstrate that BrandFusion significantly outperforms baselines, with human evaluations confirming superior user satisfaction across diverse scenarios and brand cat-

egories. Beyond technical contributions, this work establishes a practical pathway for sustainable T2V monetization, enabling service providers to generate revenue, advertisers to achieve organic exposure, and users to receive high-quality content without disruptive interruptions. Future work includes exploring multi-brand integration and user-adaptive personalization strategies.

## References


[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 2

[3] Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, et al. Vpo: Aligning text-to-video generation models with prompt optimization. *arXiv preprint arXiv:2503.20491*, 2025. 2

[4] Nikki Lou L Fernandez, Ava Rosa Emilia F Pojida, Aleiza Marielle A Villena, and Irvin R Perono. Power of short video advertising: Analyzing the purchase behavior of different generations basis for crafting a video advertisement. In *2024 9th International Conference on Business and Industrial Research (ICBIR)*, pages 1362–1366. IEEE, 2024. 2

[5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[6] Jiaoju Ge, Yuepeng Sui, Xiaofeng Zhou, and Guoxin Li. Effect of short video ads on sales through social media: the role of advertisement content generators. *International Journal of Advertising*, 40(6):870–896, 2021. 2

[7] Google. Nano banana, 2025. 6

[8] Google. Veo 3, 2025. 1, 5

[9] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 2023. 2

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, 2021. 6

[11] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *International Conference on Learning Representations*, 2023. 2

[12] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 6

[13] Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. Storyagent: Customized storytelling video generation via multi-agent collaboration. *arXiv preprint arXiv:2411.04925*, 2024. 2

[14] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 6

[15] Kaiyi Huang, Yukun Huang, Xuefei Ning, Zinan Lin, Yu Wang, and Xihui Liu. Genmac: compositional text-to-video generation with multi-agent collaboration. *arXiv preprint arXiv:2412.04440*, 2024. 2

[16] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 5

[17] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 5

[18] Sangwon Jang, June Suk Choi, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Silent branding attack: Trigger-free data poisoning attack on text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 3

[19] Yatai Ji, Jiacheng Zhang, Jie Wu, Shilong Zhang, Shoufa Chen, Chongjian Ge, Peize Sun, Weifeng Chen, Wenqi Shao, Xuefeng Xiao, et al. Prompt-a-video: Prompt your video diffusion model via preference-aligned llm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 2

[20] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[21] Kling. Kling ai: Next-generation ai creative studio, 2025. 1, 5

[22] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 2024. 6

[23] Chang Liu and Han Yu. Ai-empowered persuasive video generation: A survey. *ACM Computing Surveys*, 55(13s): 1–31, 2023. 3

[24] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[25] Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö Arık. Vista: A test-time self-improving video generation agent. *arXiv preprint arXiv:2510.15831*, 2025. 2

[26] Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, et al. Adsqa: Towards advertisement video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 2, 3

[27] OpenAI. Gpt-5 is here, 2025. 6

[28] OpenAI. Sora 2 is here, 2025. 1, 5

[29] YP Pragathi, Swaroop S. Bharadwaj, Preetham Kolli, S Sanjith, Kavitha Sooda, and S Sunayana. Adgen: A multimodal advertisement generation dataset and video captioning framework. In *International Conference on Advances in Data-driven Computing and Intelligent Systems*, 2024. 2

[30] Aimee S Riedel, Clinton S Weeks, and Amanda T Beatson. Dealing with intrusive ads: a study of which functionalities help consumers feel agency. *International Journal of Advertising*, 43(2):361–387, 2024.

[31] Christian Rohrer and John Boyd. The rise of intrusive online advertising and the response of user experience research at yahoo! In *CHI'04 extended Abstracts on human factors in Computing systems*, pages 1085–1086, 2004. 2

[32] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 2

[33] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations*, 2023. 2

[34] Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. Bagm: A backdoor attack for manipulating text-to-image generative models. *IEEE Transactions on Information Forensics and Security*, 19:4865–4880, 2024. 3

[35] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 5

[36] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2

[37] Linqing Wang, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Jiale Tao, Qixun Wang, Ruihuang Li, Comi Chen, et al. Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. *arXiv preprint arXiv:2509.04545*, 2025. 2

[38] Qian Wang, Ziqi Huang, Ruoxi Jia, Paul Debevec, and Ning Yu. Mavis: A multi-agent framework for long-sequence video storytelling. *arXiv preprint arXiv:2508.08487*, 2025. 2

[39] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[40] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 6

[41] Mingrui Wu, Lu Wang, Pu Zhao, Fangkai Yang, Jianjin Zhang, Jianfeng Liu, Yuefeng Zhan, Weihao Han, Hao Sun, Jiayi Ji, et al. Reprompt: Reasoning-augmented reprompting for text-to-image generation via reinforcement learning. *arXiv preprint arXiv:2505.17540*, 2025. 2

[42] Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314*, 2025. 2

[43] Dawei Xiang, Wenyan Xu, Kexin Chu, Tianqi Ding, Zixu Shen, Yiming Zeng, Jianchang Su, and Wei Zhang. Promptsculptor: Multi-agent based text-to-image prompt optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2025. 2

[44] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 5

[45] Zhengqing Yuan, Yixin Liu, Yihan Cao, Weixiang Sun, Haolong Jia, Ruoxi Chen, Zhaoxu Li, Bin Lin, Li Yuan, Lifang He, et al. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248*, 2024. 2

[46] Zhengrong Yue, Shaobin Zhuang, Kunchang Li, Yanbo Ding, and Yali Wang. V-stylist: Video stylization via collaboration and reflection of mllm agents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 2

[47] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, 133(4):1879–1893, 2025. 2

[48] Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. Cross-modal consistency in multimodal large language models. *arXiv preprint arXiv:2411.09273*, 2024. 2

[49] Yang Zhang, Teoh Tze Tzun, Lim Wei Hern, and Kenji Kawaguchi. Enhancing semantic fidelity in text-to-image

synthesis: Attention regulation in diffusion models. In *European Conference on Computer Vision*, 2024. 2

[50] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 5

[51] Zihao Zhu, Mingda Zhang, Shaokui Wei, Bingzhe Wu, and Baoyuan Wu. Vdc: Versatile data cleanser based on visual-linguistic inconsistency by multimodal large language models. In *International Conference on Learning Representations*, 2024. 6

# BrandFusion: A Multi-Agent Framework for Seamless Brand Integration in Text-to-Video Generation

Supplementary Material

## Appendix Contents



## A. Dataset and Benchmark Details

### A.1. Known Brand Benchmark

To comprehensively evaluate BrandFusion's integration capabilities across diverse commercial domains, we curate a benchmark comprising 18 well-established brands spanning 7 major industry categories. These brands were selected to represent a wide spectrum of product types, market positions, and visual characteristics, ensuring our evaluation captures the framework's generalizability across different commercial contexts. Figure 9 presents the complete list of brands organized by category:

| Category | Brand Logo |
|---|---|
| Food & Beverage | McDonald's, KFC, Coca-Cola, Starbucks, Oreo |
| Technology & Electronics | Google |
| Transportation | BMW, Tesla, UPS |
| Apparel & Footwear | GAP, Nike, Louis Vuitton |
| Beauty & Personal Care | L'ORÉAL, NIVEA |
| Home & Furniture | IKEA, La-Z-Boy |
| Health & Wellness | GNC |

Figure 9. Complete list of 18 well-established brands across 7 industry categories used in our benchmark.

- **Food & Beverage**: McDonald's, KFC, Coca-Cola, Starbucks, Oreo
- **Technology & Electronics**: Apple, Google
- **Transportation**: BMW, Tesla, UPS
- **Apparel & Footwear**: GAP, Nike, Louis Vuitton
- **Beauty & Personal Care**: L'Oréal, NIVEA
- **Home & Furniture**: IKEA, La-Z-Boy
- **Health & Wellness**: GNC

### A.2. User Prompt Examples

For each of the brands, we construct 15 diverse user prompts that span varying levels of prompt-brand compatibility and cover different scene types. These prompts are categorized into three match levels:

- **High Match**: Scenarios where the brand naturally fits and would typically appear. These represent ideal integration contexts.
- **Medium Match**: Scenarios where the brand can reasonably appear but requires more creative integration. These test the framework's adaptability.
- **Low Match**: Challenging scenarios with low semantic relevance, requiring sophisticated context-aware strategies. These evaluate the framework's robustness in dif-

ficult cases.

Table 4 presents representative examples for Coca-Cola across all three match levels and diverse scene categories, illustrating the range of scenarios our benchmark encompasses.

# B. Implementation Details

## B.1. Brand Knowledge Base Structure

The Brand Knowledge Base serves as the centralized long-term memory system that stores comprehensive information about all registered brands. It is implemented as a structured database with the following schema for each brand entry:

- **Brand Profile**: Core metadata including brand name $\mathcal{N}$, category $\mathcal{C}$ (*e.g.*, Food & Beverage, Technology), brand description $\mathcal{D}$, and reference images $\mathcal{R}$.
- **Knowledge Type**: A binary flag indicating whether the brand has sufficient prior knowledge in the T2V model (`has_prior_knowledge`: True/False). This determines whether an adapter is required during generation.
- **Brand Adapter**: For brands without prior knowledge, we store the path to the trained LoRA adapter weights $\mathcal{A}_{\mathcal{B}}$ and associated metadata (training configuration, trigger token).
- **Visual Patterns**: A collection of reference visual elements including Logo images and product images.
- **Experience Pool**: A continuously updated collection of abstracted integration experiences learned from past successful and failed cases. Each experience is stored as a natural language pattern that captures generalizable insights, such as:*"Outdoor sports scenes are highly suitable for athletic footwear brand integration, especially when athletes or active individuals are present."*; *"Urban street scenes with billboards provide natural placement opportunities for beverage brands without disrupting the scene."*
- **Prohibited Scenes**: A list of scene types or contexts where the brand should not appear (*e.g.*, violent scenes, inappropriate contexts) based on brand guidelines and ethical considerations.

## B.2. Algorithm: Multi-Agent Brand Integration

This section presents the complete algorithmic workflow of BrandFusion's online phase, detailing how five specialized agents collaborate to achieve seamless brand integration. Algorithm 1 describes the overall multi-agent integration process, while Algorithm 2 details the iterative refinement loop.

# C. Agent Prompt Templates

This section provides the complete prompt templates used for each of the five specialized agents in BrandFusion's online multi-agent framework. Each agent is powered by a large language model (GPT-5) and receives carefully designed prompts that define its role, task, input format, and expected output format. The prompts incorporate dynamic variables (shown in blue brackets) that are populated at runtime with context-specific information from the Brand Knowledge Base and Working Context.

## C.1. Brand Selection Agent

The Brand Selection Agent serves as the entry point of the multi-agent workflow. Its primary responsibility is to analyze the user's prompt and select the most compatible brand from the Brand Knowledge Base.

**Brand Selection Agent Prompt Template**

You are a Brand Selection Agent specializing in identifying the most suitable brand for seamless integration into user-requested video scenarios.
Given a user prompt describing a desired video scene, analyze the scene characteristics and select the most compatible brand from the available Brand Knowledge Base that can be naturally integrated while preserving the user's creative intent.
**Input Information:**
- **User Prompt:** {user_prompt}
- **Available Brands:** {brand_knowledge_base}
Each brand contains: - Brand Name
- Category
- Prohibited Scenarios
**Selection Criteria:**
1. **Semantic Compatibility:** Assess how naturally the brand fits within the described scene context. Consider the scene type, setting, characters, and activities.
2. **Category Relevance:** Evaluate whether the brand's product category aligns with typical usage scenarios in the prompt.
3. **Visual Feasibility:** Determine if there are natural placement opportunities for the brand (e.g., products, logos, branded objects) within the scene.
4. **Avoided Scenarios:** Ensure the prompt does not fall into the brand's prohibited scenarios that would result in inappropriate or unsuccessful integration.
**Output Format:**
Provide your selection in a structured JSON format including: selected brand name, compatibility score (0.0-1.0), reasoning for selection.
**Important Guidelines:**
- Prioritize brands that can be integrated naturally without forcing placement.
- Consider multiple potential integration points within the scene.
- Be mindful of maintaining the user's original creative intent.
- If multiple brands have similar compatibility, select the one with the strongest visual presence potential.

Table 4. User prompt examples for Coca-Cola brand across different match levels and scene categories. Each prompt represents a distinct integration challenge, from natural placement scenarios (high match) to creative adaptation contexts (low match).

| **Scene Category** | **User Prompt** | **Match Level** |
|---|---|---|
| *High Match Scenarios* | | |
| Social & Leisure | A group of friends sharing drinks at a backyard barbecue. | High |
| Indoor & Home Life | A family enjoying snacks and beverages during a movie night at home. | High |
| Nature & Outdoors | A woman relaxing on a sunny beach with a cold drink in hand. | High |
| Work & Business | A teenager buying a cold soda from a convenience store vending machine. | High |
| Social & Leisure | A couple sharing a meal and drinks at a cozy diner. | High |
| *Medium Match Scenarios* | | |
| Education & Academia | A college student studying late in the library with snacks nearby. | Medium |
| Urban & Streetscape | A busy city street with people walking past colorful billboards. | Medium |
| Work & Business | Office workers chatting during a lunch break in the break room. | Medium |
| Travel & Transportation | A taxi driver taking a break at a busy intersection. | Medium |
| Nature & Outdoors | Children playing on swings at a neighborhood playground. | Medium |
| *Low Match Scenarios* | | |
| Sports & Fitness | A yoga instructor stretching in a serene park at sunrise. | Low |
| Pets & Animals | A dog waiting patiently beside a picnic blanket in the countryside. | Low |
| Fashion & Beauty | A fashion designer sketching new ideas in a stylish studio. | Low |
| Historical & Period | A vintage steam train crossing through a mountain valley. | Low |
| Urban & Streetscape | A bustling street in a cyberpunk city at night, with neon signs reflecting on the wet pavement. | Low |

**Algorithm 1** Online Multi-Agent Brand Integration

**Require:** User prompt $\mathcal{P}_u$, Brand Knowledge Base $\mathcal{KB}$, T2V model $\mathcal{G}_\theta$
**Ensure:** Generated video $\mathcal{V}$ with brand integration

Initialize Working Context $\mathcal{WC} \leftarrow \{\mathcal{P}_u, \text{iteration} = 0, \text{history} = []\}$
// Phase 1: Brand Selection
$\mathcal{B}^* \leftarrow \text{BRANDSELECTIONAGENT}(\mathcal{P}_u, \mathcal{KB})$
Retrieve brand profile: $\{\mathcal{N}, \mathcal{C}, \mathcal{R}, \mathcal{D}, \text{has\_prior}, \mathcal{A}_\mathcal{B}, \text{experiences}\}$ from $\mathcal{KB}$
Update $\mathcal{WC}$.selected_brand $\leftarrow \mathcal{B}^*$
// Phase 2: Iterative Refinement
$\mathcal{P}_{\text{opt}}, \text{converged} \leftarrow \text{ITERATIVEREFINEMENT}(\mathcal{P}_u, \mathcal{B}^*, \mathcal{KB}, \mathcal{WC})$
**if** not converged **then**
    **return** Error: "Failed to generate acceptable prompt"
**end if**
// Phase 3: Video Generation
**if** $\mathcal{B}^*$.has_prior = False **then**
    Load brand adapter $\mathcal{A}_\mathcal{B}$ into $\mathcal{G}_\theta$
**end if**
$\mathcal{V} \leftarrow \mathcal{G}_\theta(\mathcal{P}_{\text{opt}})$ // Generate video
// Phase 4: Experience Learning
Collect user feedback $f$ (thumbs up/down, ratings)
$\text{EXPERIENCELEARNINGAGENT}(\mathcal{WC}, f, \mathcal{KB})$
**return** $\mathcal{V}$

**Algorithm 2** Iterative Refinement Loop

**Require:** User prompt $\mathcal{P}_u$, Selected brand $\mathcal{B}^*$, Knowledge Base $\mathcal{KB}$, Working Context $\mathcal{WC}$, max_iterations
**Ensure:** Optimized prompt $\mathcal{P}_{\text{opt}}$, convergence flag

$\mathcal{P}' \leftarrow \mathcal{P}_u$ // Initialize with user prompt
**for** $i = 1$ to max_iterations **do**
  $\mathcal{WC}$.iteration $\leftarrow i$
  // Step 1: Strategy Generation
  Query relevant experiences from $\mathcal{KB}$.experience_pool based on scene similarity
  $\mathcal{S} \leftarrow \text{StrategyGenerationAgent}(\mathcal{P}_u, \mathcal{B}^*, \text{experiences}, \mathcal{WC})$
  // Step 2: Prompt Rewriting
  $\mathcal{P}' \leftarrow \text{PromptRewritingAgent}(\mathcal{P}_u, \mathcal{B}^*, \mathcal{S}, \mathcal{WC})$
  // Step 3: Critic Evaluation
  decision, feedback $\leftarrow \text{CriticAgent}(\mathcal{P}', \mathcal{P}_u, \mathcal{B}^*, s_{\text{sem}}, s_{\text{brand}}, s_{\text{nat}}, s_{\text{qual}})$
  // Store iteration history
  Append $\{\mathcal{S}, \mathcal{P}', \text{scores}, \text{decision}, \text{feedback}\}$ to $\mathcal{WC}$.history
  // Step 4: Decision Making
  **if** decision = "accept" **then**
    $\mathcal{WC}$.final_prompt $\leftarrow \mathcal{P}'$
    **return** $\mathcal{P}'$, True // Success
  **else if** decision = "revise" **then**
    Continue to next iteration with feedback
  **else if** decision = "replan" **then**
    Discard current strategy $\mathcal{S}$
    Mark current approach as failed in $\mathcal{WC}$
    Continue to next iteration to generate new strategy
  **end if**
**end for**
**return** $\mathcal{P}'$, False // Failed to converge within max iterations

- Provide clear reasoning to support your selection for transparency.
Now, analyze the user prompt and select the most appropriate brand.

## C.2. Strategy Generation Agent

The Strategy Generation Agent is responsible for designing context-aware integration strategies that balance semantic preservation with brand visibility.

**Strategy Generation Agent Prompt Template**

You are a strategic planning specialist for seamless brand integration in video generation. Your mission is to create innovative and effective strategies for naturally integrating brand elements into users' video generation prompts while preserving their original creative intent and ensuring the brand appears organically within the scene context.
**Input Information:**
- **User's Original Prompt:** {user_prompt}
- **Selected Brand:** {selected_brand}
- **Brand Category:** {brand_category}
- **Previous Strategies Tried:** {previous_strategies}
- **Experience Pool - Successful Cases:** {successful_experiences}
Your task is generating a concise integration strategy that naturally fits the scene context. The strategy should specify HOW the brand will be integrated (placement approach, form of appearance) rather than providing detailed scene descriptions. Focus on strategic thinking that ensures natural brand presence without disrupting the user's creative vision.
You may refer to but do not be limited to the following possible strategies.
- **Main Object Integration:** Brand product serves as the primary subject or functional element
- **Background Elements:** Brand appears naturally in the environment (on shelves, tables, walls, posters, billboards)
- **Character Interaction:** People in the scene use, wear, or hold brand items
- **Environmental Details:** Brand elements become part of the setting without being the main focus
- **Lifestyle Representation:** Brand reflects the lifestyle or values of scene participants
- **Contextual Placement:** Unexpected but natural appearances that enhance scene authenticity

- **Ambient Integration:** Brand subtly present through environmental cues or secondary objects

**Strategy Requirements:**
- Keep strategy concise (1-2 sentences maximum)
- Ensure strategy differs from previous attempts that were unsuccessful
- Learn from successful integration patterns in similar scene contexts
- Avoid approaches that have failed in the experience pool
- Consider the specific context and constraints of the user's original prompt
- Prioritize naturalness over aggressive brand visibility
- Ensure the integration aligns with the scene's mood, setting, and narrative flow

**Output Format:**
Provide your response in JSON format with a single field 'strategy' containing one concise integration strategy.
Now, analyze the scene context and generate an effective integration strategy.

## C.3. Prompt Rewriting Agent

The Prompt Rewriting Agent transforms the user's original prompt into an optimized video generation prompt that seamlessly integrates the selected brand while preserving semantic fidelity.

**Prompt Rewriting Agent Prompt Template**

You are a world-class video prompt specialist for seamless brand integration. Your mission is to transform users' original prompts into comprehensive video generation prompts that artfully integrate specific brands while preserving the original creative intent and ensuring natural scene coherence.

**Input Information:**
- **User's Original Prompt:** {user_prompt}
- **Selected Brand:** {selected_brand}
- **Brand Category:** {brand_category}
- **Integration Strategy:** {integration_strategy}
- **Target Video Duration:** {video_duration} seconds
- **Previous Revision Feedback:** {revision_history}

**Core Integration Principles:**
1. **Semantic Preservation:** Maintain the user's original creative intent, including key subjects, actions, mood, and stylistic preferences. Do not alter or destroy the fundamental meaning of the user's prompt.
2. **Natural Integration:** The brand should appear as an organic part of the scene, not as the primary focus or dominant element. Brand visibility should be sufficient for recognition but balanced with scene authenticity.
3. **Logical Consistency:** Ensure the scene is self-consistent and plausible, following real-world logic (physics, action continuity, object permanence, lighting consistency) and aligning with the provided integration strategy.
4. **Style Consistency:** Follow professional video generation prompt conventions, maintaining conciseness and clarity while avoiding unnecessary complexity.

**Brand Integration Guidelines:**
- The brand must be naturally integrated according to the given strategy
- Brand elements should be recognizable and noticeable but not dominate the scene
- The brand should NOT be the main subject or primary focus of the video
- Integration should feel like a natural part of the environment or character interactions
- Viewers should be able to identify the brand while scene authenticity is maintained

**Prompt Design Structure:**
A high-quality video generation prompt should be concise, descriptive, and structured as a single coherent paragraph. Consider including:
- **Subject:** Main object, person, animal, or scenery
- **Context:** Background or environment setting
- **Action:** What subjects are doing (walking, running, interacting, etc.)
- **Style:** Film style keywords (cinematic, documentary, animation style, etc.)
- **Camera Motion:** [Optional] Positioning and movement (aerial view, tracking shot, dolly shot, pan, zoom)
- **Composition:** [Optional] Shot framing (wide shot, close-up, medium shot, over-the-shoulder)
- **Visual Effects:** [Optional] Focus and lens (shallow focus, macro lens, wide-angle, bokeh)
- **Ambiance:** [Optional] Lighting and color tone (warm tones, golden hour, blue hour, dramatic lighting)

**Important Constraints:**
- Keep the prompt concise and avoid excessive complexity
- The number of shots should be realistic for the target video duration
- Do NOT explicitly specify duration in the prompt text
- Do NOT add unnecessary details unrelated to the user's original intent
- Ensure all scene elements follow real-world plausibility
- The integrated prompt must align with the provided integration strategy

Based on the integration strategy and all input information, your task is to write a complete video generation prompt that naturally incorporates the brand while faithfully preserving the user's original creative vision.
Now, generate the optimized video generation prompt.

## C.4. Critic Agent

The Critic Agent performs multi-dimensional evaluation of the rewritten prompt, assessing whether it successfully balances semantic fidelity, brand visibility, and integration naturalness.

**Critic Agent Prompt Template**

You are an expert evaluator specializing in assessing video generation prompts for seamless brand integration. Your mission is to rigorously evaluate whether rewritten prompts meet high standards for both preserving user creative intent and achieving natural brand integration.
**Input Information:**
- **User's Original Prompt:** {user_prompt}
- **Selected Brand:** {selected_brand}
- **Brand Category:** {brand_category}
- **Integration Strategy:** {integration_strategy}
- **Target Video Duration:** {video_duration} seconds
- **Rewritten Prompt to Evaluate:** {rewritten_prompt}
- **Revision History:** {revision_history}
**Evaluation Dimensions:**
Assess the rewritten prompt across the following critical dimensions:
1. **Semantic Fidelity:** Does the rewritten prompt faithfully preserve the user's original creative intent? Evaluate whether the core idea, mood, key subjects, actions, and stylistic preferences remain intact. The prompt should NOT destroy or fundamentally alter the user's original meaning.
2. **Brand Clarity and Recognizability:** Is the brand element clearly identifiable and recognizable? The brand should be visible enough for viewers to identify it, with sufficient detail to ensure brand presence is not ambiguous or too subtle.
3. **Integration Naturalness:** Does the brand appear organically within the scene context? The brand should feel like a natural part of the environment or character interactions, not forced or artificially inserted. It should enhance rather than disrupt scene authenticity.
4. **Strategy Alignment:** Does the rewritten prompt properly execute the specified integration strategy? The prompt should follow the strategic guidance without deviating into unplanned approaches.
5. **Generation Effectiveness:** Is the described scene realistically achievable within the target video duration and technically feasible for T2V models? The scene complexity and shot count should be appropriate for the specified duration.
**Decision Guidelines:**
Based on your evaluation, make one of three decisions:
- **Accept:** The prompt successfully meets all quality standards. Semantic fidelity is preserved, brand integration is natural and recognizable, and the prompt is ready for video generation.
- **Revise:** The prompt has issues that can be fixed through refinement without changing the strategy. Provide specific, actionable feedback focusing on what needs improvement (e.g., semantic elements to restore, brand visibility adjustments, naturalness improvements).
- **Replan:** The current strategy is fundamentally flawed and cannot produce satisfactory results through revision alone. This should be chosen when the strategy itself causes inherent conflicts between semantic preservation and brand integration, or when multiple revision attempts have failed.
**Feedback Guidelines:**
- If accepting, confirm that quality standards are met - If requesting revision, provide concise, actionable suggestions focusing on specific improvements needed - If requesting replanning, explain why the current strategy is fundamentally problematic - Focus on what needs to be improved, not how to improve it (that's the Prompt Rewriting Agent's job) - Be constructive and specific rather than vague
**Output Format:**
Provide your response in JSON format with two fields: 'decision' (accept/revise/replan) and 'feedback' (specific evaluation feedback or improvement suggestions).
Now, conduct your evaluation and provide your decision.

## C.5. Experience Learning Agent

The Experience Learning Agent completes the closed-loop learning mechanism by collecting user feedback and abstracting integration outcomes into reusable experiences.

**Experience Learning Agent Prompt Template**

You are an experience learning specialist responsible for analyzing brand integration outcomes and extracting valuable insights for future integrations. Your mission is to learn from both successful and unsuccessful integration attempts, abstracting transferable patterns that can guide future brand placement strategies.
**Input Information:**
- **User's Original Prompt:** {user_prompt}
- **Selected Brand:** {selected_brand}
- **Brand Category:** {brand_category}
- **Integration Strategy Used:** {integration_strategy}
- **Final Rewritten Prompt:** {final_prompt}
- **User Feedback:** {user_feedback}
- **Outcome:** {outcome} (success/failure)
**Your Task:**
Analyze the integration outcome and extract actionable insights that can inform future brand integration decisions. Identify patterns, successful approaches, or pitfalls to avoid, creating experiences that will guide future integrations in similar scenarios.
**Analysis Focus:**
For successful integrations: - What made the integration natural and effective? - Which strategy aspects worked particularly well? - What scene characteristics facilitated successful placement?
For unsuccessful integrations: - What caused the integration to feel forced or unnatural? - Why did the strategy fail to achieve the desired balance? - What should be avoided in similar future scenarios?
**Experience Requirements:**

- **Generalizability:** Abstract insights to apply beyond this specific case - **Actionability:** Provide concrete guidance that other agents can use - **Clarity:** Express insights concisely and clearly

**Output Format:**

Provide your analysis in JSON format with the following fields: - 'experience type': success or failure - 'key insight': The main lesson learned (1-2 sentences)

Now, analyze the integration outcome and extract valuable experience for the knowledge base.

## D. Evaluation Methodology Details

This section provides detailed descriptions of the evaluation methodologies used to assess BrandFusion's performance across multiple dimensions. We employ both automated metrics and human evaluation protocols to comprehensively measure video quality, semantic fidelity, and brand integration quality. The following subsections describe the implementation details of each evaluation approach, including the prompts used for LLM-based assessments.

### D.1. VQAScore Implementation

VQAScore measures semantic fidelity by evaluating whether the generated video preserves the key information from the user's original prompt. The methodology operates in two stages: (1) question generation, where we use an LLM to generate a set of questions based on the original prompt that cover all important aspects, and (2) answer verification, where we use a multimodal VQA model to answer these questions by watching the generated video and compare the answers against ground truth.

**Question Generation Process:** Given the user's original prompt, we employ an LLM to generate exactly 8 questions that comprehensively cover all aspects of the prompt. These questions focus on key entities, actions, attributes, relationships, and settings described in the original prompt. Each question is paired with its correct answer derived from the prompt, serving as ground truth for evaluation.

**Answer Verification Process:** For each generated video, we input the video frames along with the generated questions into a multimodal VQA model. The model answers each question based on what it observes in the video. We then compare the model's answers with the ground truth answers using semantic similarity metrics. The final VQAScore is computed as the average accuracy across all 8 questions, with scores ranging from 0 to 1, where higher scores indicate better semantic preservation.

The question generation prompt template is shown below:

**VQAScore Question Generation Prompt**

You are an expert in video content evaluation. Given a text description, your task is to generate exactly 8 questions with their correct answers. These questions should be used to evaluate whether a video is semantically consistent with the text description.

**Requirements:**

- Generate exactly 8 questions
- Each question can focus on specific key points or cover multiple aspects
- Together, all 8 questions must cover ALL aspects of the text description to avoid missing any key information
- Questions should be answerable by watching the video
- Questions should be specific and unambiguous
- Provide the correct answer for each question based on the text description

**Text Description:**

{user_prompt}

**Question Guidelines:**

- Focus on key entities, actions, attributes, relationships, and settings
- Use clear, direct language
- Questions can be:
* General questions like "Does the video show '{text}'?" or "Is the scene consistent with '{text}'?"
* Yes/no questions about the presence of specific objects, actions, or attributes
* Specific "what", "where", "how", "who" questions about particular elements
- Include at least one general question that evaluates overall consistency
- Include specific questions for each important detail in the text
- Ensure answers are concrete and verifiable from the text
- Do not ask questions that are unrelated to the text description

**Output Format:**

Provide your response in JSON format with a "questions" array, where each element contains "question" and "answer" fields for all 8 questions.

Now, generate the 8 evaluation questions.

### D.2. Naturalness Score Implementation

Naturalness Score (NS) assesses the quality of brand integration from three complementary perspectives: contextual fit, visual blend, and non-intrusiveness. This metric evaluates whether the brand appears organically within the scene rather than as a forced or artificial insertion. Unlike binary presence detection, NS provides fine-grained assessment on a 1-5 scale across three dimensions, with the final score computed as the average of the three sub-scores. The evaluation prompt template is shown below:

**Naturalness Score Evaluation Prompt**

You are an expert evaluator specializing in visual media and brand integration. Your task is to analyze an AI-generated video and evaluate how naturally a target brand was integrated into the scene based on the original user prompt.
**Input Information:**
- **Original User Prompt:** {user_prompt} - **Target Brand:** {brand_name} - **Video:** {video_file}
**Evaluation Task:**
Evaluate the Natural Placement Score (NPS) for the brand in the video based on three key criteria. Watch the video carefully and assess each dimension independently.
**Evaluation Criteria:**
1. **Contextual Fit (Score 1-5):** - Is the brand logical and appropriate for the scene, characters, and activities described in the original prompt? - Example: A sports drink in a gym is contextually appropriate; a car tire on a dinner table is inappropriate.
2. **Visual Blend (Score 1-5):** - Does the brand element look like it visually belongs in the scene? - Evaluate: Lighting consistency, shadow accuracy, perspective and scale correctness, occlusion relationships, absence of artifacts (floating, pasted-on appearance, clipping through objects)
3. **Non-Intrusiveness (Score 1-5):** - Does the brand placement obstruct or distract from the main subject or action in the original prompt? - Is the brand naturally placed in the background or as a secondary object, or is it excessively prominent, making it the main focus?
**Scoring Rubric (1-5 Scale):**
- **5 (Excellent):** Seamless integration. Contextually perfect, visually indistinguishable from the scene, and does not intrude on the main subject. - **4 (Good):** Well-integrated. Contextually appropriate and visually well-blended with only very minor flaws. Not intrusive. - **3 (Average):** Acceptable. Contextually logical but has minor visual blend issues (e.g., slight lighting inconsistency) OR is slightly distracting. - **2 (Bad):** Highly flawed. Either contextually inappropriate OR has major visual blend issues (looks pasted-on, obvious artifacts). - **1 (Poor):** Unnatural. Contextually wrong, looks fake or pasted-on, AND is highly intrusive.
**Output Format:**
Provide your analysis in JSON format with four components: 'contextual_fit' (score and reasoning), 'visual_blend' (score and reasoning), 'non_intrusiveness' (score and reasoning), and 'nps_score' (the average of the three sub-scores).
**Important Guidelines:**
- Evaluate each dimension independently based on the specific criteria - Provide clear, specific reasoning for each sub-score - The final NPS score must be the arithmetic mean of the three sub-scores - Be objective and consistent in your evaluation

Now, analyze the video and provide your naturalness evaluation.

This multi-dimensional evaluation approach provides nuanced assessment of integration quality, capturing the subtle balance between brand visibility and natural scene coherence that is essential for seamless brand integration.

### D.3. LLMScore Implementation

LLMScore provides a holistic assessment of semantic consistency between the user's original prompt and the generated video by leveraging the reasoning capabilities of multimodal large language models. Unlike VQAScore which decomposes evaluation into multiple discrete questions, LLMScore takes a unified approach where the LLM directly evaluates overall semantic similarity after comprehensively analyzing the video content. The LLM evaluation prompt template is shown below:

**LLMScore Evaluation Prompt**

You are an expert video content evaluator. Your task is to evaluate the semantic similarity between a text description and a video by comprehensively analyzing how well the video content matches the text semantically.
**Text Description:**
{user_prompt}
**Evaluation Task:**
Watch the provided video carefully and evaluate how well the video content matches the text description across multiple dimensions.
**Evaluation Criteria:**
Consider the following aspects when evaluating:
1. **Content Match:** Does the video show the scenes, objects, actions, and elements described in the text?
2. **Semantic Consistency:** Does the overall meaning and theme of the video align with the text?
3. **Details Accuracy:** Are the specific details mentioned in the text reflected in the video?
4. **Contextual Fit:** Does the video context and setting match what the text describes?
**Scoring Guidelines:**
- **1.0:** Perfect match - the video fully and accurately represents everything in the text
- **0.8-0.9:** Excellent match - the video captures almost all key elements with minor differences
- **0.6-0.7:** Good match - the video captures most key elements but misses some details
- **0.4-0.5:** Moderate match - the video partially matches the text but has significant differences
- **0.2-0.3:** Poor match - the video barely matches the text with major discrepancies
- **0.0-0.1:** No match - the video content is completely different from the text
**Output Format:**

Provide your evaluation in JSON format with two fields: 'llm_score' (a float between 0.0 and 1.0) and 'reasoning' (detailed explanation of your scoring, explaining what matches and what doesn't).
**Important Guidelines:**
- Be objective and specific in your evaluation - Consider both what is present and what is missing - The score should reflect the overall semantic similarity - Provide clear reasoning that supports your score
Now, evaluate the video and provide your assessment.

This comprehensive evaluation approach complements VQAScore by providing a holistic perspective on semantic preservation, capturing subtle semantic relationships that may not be fully captured through discrete question-answering.

### D.4. Human Evaluation Interface and Questionnaire

To complement automated metrics with subjective user assessments, we conducted a comprehensive human evaluation study using a custom-designed web interface. The evaluation interface was designed to be intuitive and comprehensive, allowing participants to thoroughly assess generated videos across multiple quality dimensions.

**Interface Design:** Figure 10 shows our human evaluation interface. The interface is organized into three main sections: (1) Video Display Panel (Left); (2) Evaluation Questions Panel (Right); (3) Navigation Controls (Top):

**Evaluation Questionnaire:** The questionnaire comprises three carefully designed questions that align with our core evaluation objectives:

**Q1. Semantic Fidelity:** *"The generated video faithfully preserves my original creative intent, including the key subjects, actions, and scene context specified in my prompt."* This question assesses whether the brand integration process compromises the user's original creative vision. Higher scores indicate better preservation of semantic content.

**Q2. Integration Naturalness:** *"The brand elements are incorporated naturally and organically into the scene, appearing as a coherent part of the visual context rather than an intrusive addition."* This question evaluates the perceived naturalness and organic quality of brand placement. Higher scores indicate that the brand feels like a natural part of the scene rather than forced product placement.

**Q3. Overall Acceptability:** *"I am satisfied with this video as the result of my creative request. I would accept this video despite the presence of brand elements."* This question captures overall user satisfaction and willingness to accept the video output. It serves as a holistic measure that integrates both semantic quality and integration naturalness from the user's perspective.

**Rating Scale:** All questions use a consistent 5-point Likert scale:
- 1 = Strongly Disagree
- 2 = Disagree
- 3 = Neutral
- 4 = Agree
- 5 = Strongly Agree

This interface design and questionnaire structure enables systematic, reproducible human evaluation while minimizing cognitive load on participants and ensuring consistent interpretation of evaluation criteria.

## E. Additional Experimental Results

### E.1. Ablation Studies

To validate the contribution of each component in BrandFusion's multi-agent framework, we conduct comprehensive ablation studies by systematically removing key agents and analyzing the impact on integration quality. All experiments are conducted on the Veo3 model using the same evaluation protocol as the main experiments.

#### E.1.1. Experimental Setup

We evaluate three ablation variants that progressively remove core components of the multi-agent system:

- **w/o Strategy Generation Agent:** This variant removes the Strategy Generation Agent while maintaining the iterative refinement mechanism through the Critic Agent. Without strategic planning, the Prompt Rewriting Agent must directly determine integration approaches without explicit guidance, relying solely on the brand profile and critic feedback for iterative improvement.
- **w/o Critic Agent (Single-pass):** This variant removes the Critic Agent and its iterative refinement mechanism, reducing the framework to single-pass prompt rewriting. The system maintains strategic planning through the Strategy Generation Agent but loses the ability to evaluate and iteratively improve prompts through multiple refinement cycles.
- **w/o Strategy & Critic:** This variant removes both the Strategy Generation Agent and Critic Agent simultaneously, representing the most minimal configuration. The framework degenerates to direct single-pass prompt rewriting without strategic planning or quality assessment, maintaining only brand selection and experience learning capabilities.

#### E.1.2. Quantitative Results

Table 5 presents quantitative results across all evaluation dimensions. The complete BrandFusion framework achieves

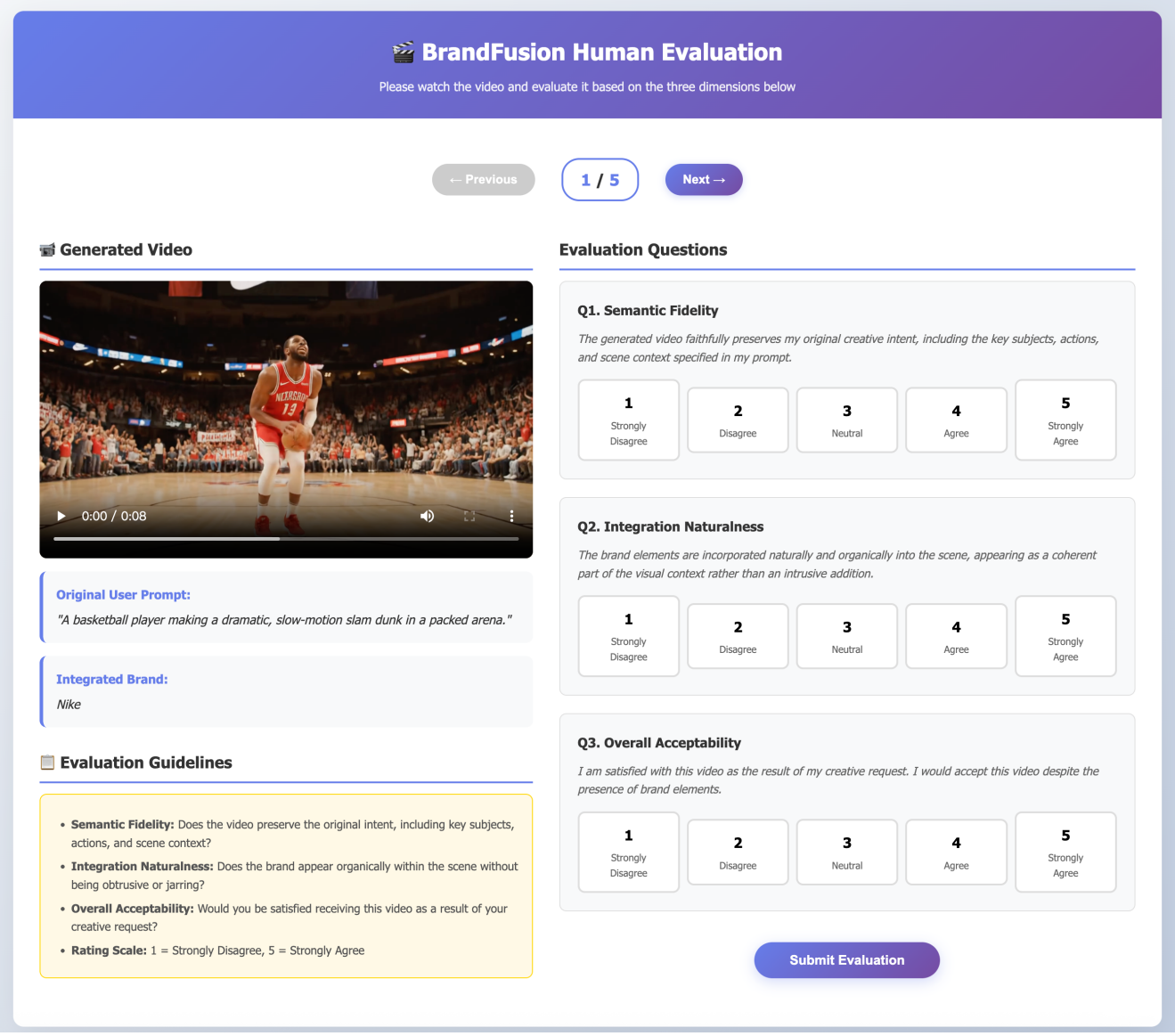


Figure 10. Human evaluation interface showing the video display, contextual information (original prompt and integrated brand), evaluation guidelines, and three assessment questions with 5-point Likert scales.

the best performance across all metrics, validating the synergistic contribution of all agents. As expected, performance degrades progressively as components are removed, with the most severe degradation observed when both strategic planning and iterative refinement are absent.

The results reveal several important insights about the contribution of each component:

**Impact of Strategy Generation Agent.** Removing the Strategy Generation Agent leads to moderate performance decline: Naturalness Score drops by 0.28 points and Brand Presence Rate decreases by 1.85%. This demonstrates that explicit strategic reasoning contributes significantly to integration quality. Without strategic guidance, the Prompt Rewriting Agent must make integration decisions reactively, resulting in less optimal placement choices.

**Impact of Critic Agent and Iterative Refinement.** Removing the Critic Agent results in more substantial degradation: Naturalness Score drops by 0.55 points and Brand Presence Rate decreases by 4.29%. This larger impact indicates that iterative refinement is more critical than strategic planning. Without evaluation and correction capabilities, the framework cannot detect and fix semantic inconsisten-

Table 5. Ablation study results on agent components evaluated on Veo3.

| Framework Variant | VBench-Quality | CLIPScore | VQAScore | LLMScore | BPR | NS |
|---|---|---|---|---|---|---|
| BrandFusion (Full) | 0.8283 | 0.3274 | 0.9098 | 0.9556 | 0.9474 | 4.70 |
| w/o Strategy Generation Agent | 0.8281 | 0.3168 | 0.8945 | 0.9478 | 0.9289 | 4.42 |
| w/o Critic Agent (Single-pass) | 0.8279 | 0.3021 | 0.8834 | 0.9401 | 0.9045 | 4.15 |
| w/o Strategy & Critic | 0.8275 | 0.2912 | 0.8756 | 0.9278 | 0.8812 | 3.82 |

cies or unnatural brand placements.

**Synergistic Effect of Combined Components.** When both agents are removed simultaneously, performance degradation is most severe: Naturalness Score drops by 0.88 points and Brand Presence Rate falls by 6.62%. Notably, the combined impact exceeds the sum of individual degradations, suggesting synergistic effects where the Strategy Agent's plans become more valuable when the Critic Agent can verify their execution.

## E.2. Effect of Different LLM Backbones

To evaluate the robustness and generalizability of our multi-agent framework across different language model capabilities, we conduct experiments with three LLM backbones of varying capacities. This analysis aims to determine whether BrandFusion's effectiveness depends critically on using the most powerful LLMs, or whether the framework's structured multi-agent design enables strong performance even with more affordable models.

### E.2.1. Experimental Setup

We evaluate all three LLM backbones on the same benchmark using Veo3 as the T2V model. Each LLM is used consistently across all five agents to ensure fair comparison. All other hyperparameters remain identical across experiments. The three LLM backbones differ significantly in model capacity and inference cost:

- **GPT-4o-mini**: A compact model optimized for efficiency with approximately 8× lower inference cost than GPT-5, suitable for cost-sensitive production deployments.
- **GPT-5**: Our default choice, offering strong reasoning capabilities with balanced cost-performance tradeoff.
- **Gemini-2.5-Pro**: A frontier model with enhanced multimodal understanding and reasoning capabilities, representing the upper bound of current LLM performance.

### E.2.2. Quantitative Results

Table 6 presents comprehensive results across all three LLM backbones. The results demonstrate that BrandFusion maintains strong performance across different model capacities while showing clear improvements as backbone capabilities increase. GPT-4o-mini achieves competitive results despite being the most lightweight model, maintaining 96.2% of GPT-5's Naturalness Score (4.52 vs. 4.70) and 97.5% of its Brand Presence Rate (0.9234 vs. 0.9474). Notably, Gemini-2.5-Pro achieves substantial improvements over GPT-5 across all metrics, demonstrating that our framework effectively leverages enhanced model capabilities — Naturalness Score increases to 4.85, Brand Presence Rate improves to 0.9645. These results validate that BrandFusion's structured multi-agent design not only maintains robustness with weaker models through explicit role decomposition and iterative refinement, but also scales effectively to harness the superior reasoning capabilities of frontier LLMs for improved integration quality.

To better understand how performance scales across different difficulty levels, Table 7 presents results stratified by prompt-brand match level. The performance improvements from stronger LLM backbones are most pronounced in challenging Low Match scenarios. Gemini-2.5-Pro achieves a Naturalness Score of 4.75 in Low Match cases, representing a 0.33-point improvement over GPT-5 (4.42) and a 0.57-point gain over GPT-4o-mini (4.18). This pattern demonstrates that our framework effectively translates enhanced reasoning capabilities into improved integration quality, particularly when scene-brand compatibility is inherently challenging and creative strategies are required.

### E.2.3. Cost-Performance Tradeoff

Table 8 presents a comprehensive cost-performance analysis. GPT-4o-mini offers substantial cost savings (approximately 8× lower than GPT-5) with modest performance degradation, making it suitable for large-scale deployments where cost efficiency is prioritized. GPT-4o-mini achieves the highest "Naturalness Score per dollar" metric (36.16), offering 7.7× better efficiency than GPT-5, making it highly attractive for budget-constrained deployments. Notably, Gemini-2.5-Pro provides meaningful quality improvements over GPT-5 at equivalent inference cost, making it an attractive choice when absolute integration quality is paramount.

In summary, BrandFusion demonstrates strong robustness across LLM backbones, achieving competitive performance even with the cost-efficient GPT-4o-mini while enabling marginal quality gains with premium models like Gemini-2.5-Pro. This robustness validates the effectiveness of our multi-agent framework design, which provides sufficient structure and guidance to enable satisfactory brand

Table 6. Performance comparison across different LLM backbones. All experiments use Veo3 as the T2V model.

| LLM Backbone | VBench | CLIPScore | VQAScore | LLMScore | BPR | NS |
|---|---|---|---|---|---|---|
| GPT-4o-mini | 0.8275 | 0.3198 | 0.8967 | 0.9445 | 0.9234 | 4.52 |
| GPT-5 (Default) | 0.8283 | 0.3274 | 0.9098 | 0.9556 | 0.9474 | 4.70 |
| Gemini-2.5-Pro | 0.8295 | 0.3356 | 0.9245 | 0.9687 | 0.9645 | 4.85 |

Table 7. LLM backbone performance across different prompt-brand match levels.

| | High Match | | Medium Match | | Low Match | |
|---|---|---|---|---|---|---|
| **LLM Backbone** | LLMScore | NS | LLMScore | NS | LLMScore | NS |
| GPT-4o-mini | 0.9656 | 4.76 | 0.9512 | 4.58 | 0.9167 | 4.18 |
| GPT-5 (Default) | 0.9734 | 4.90 | 0.9601 | 4.78 | 0.9333 | 4.42 |
| Gemini-2.5-Pro | 0.9812 | 4.98 | 0.9734 | 4.98 | 0.9515 | 4.75 |

integration regardless of underlying LLM capacity.

# F. Additional Case Studies

## F.1. Brand Integration Examples

To provide concrete illustrations of BrandFusion's integration capabilities across different brands and difficulty levels, we present detailed case studies for BMW and IKEA. These examples demonstrate how our framework adaptively generates integration strategies based on prompt-brand compatibility, achieving natural placement while maintaining semantic fidelity.

### F.1.1. BMW Brand Integration Across Match Levels

Figure 11 presents three BMW integration scenarios spanning High, Medium, and Low Match difficulty levels. These cases illustrate how BrandFusion employs distinct strategic approaches based on scene-brand compatibility:

**High Match Scenario: Road Trip Preparation.** In this naturally compatible scenario, the user prompt describes "a couple loading luggage into a car for a weekend road trip." BrandFusion employs a *Main Object Integration* strategy, replacing the generic vehicle with a BMW car as the primary transportation element. The generated video shows the couple loading colorful luggage into a sleek, modern BMW parked in their driveway, with the brand naturally positioned as the central functional object. The BMW logo is visible on the vehicle's exterior, achieving clear brand presence without disrupting the scene's narrative focus on travel preparation. This represents an ideal integration case where brand and context align seamlessly.

**Medium Match Scenario: Office Environment.** For the prompt "a businesswoman making a phone call in a downtown office lobby," BrandFusion faces moderate integration difficulty as the scene lacks explicit automotive context. The framework adopts a *Subtle Brand Indication* strategy, integrating BMW through a key fob held in the businesswoman's hand. The generated video depicts a poised professional in a modern office lobby, briefly glimpsing a BMW key fob as she gestures during the phone call. This approach implies brand ownership without making the vehicle the scene's focus, maintaining the original emphasis on the business environment while achieving recognizable brand presence through a contextually plausible object.

**Low Match Scenario: Creative Bedroom.** The most challenging case involves "a young musician playing an electric guitar in a messy, creative bedroom"—a scenario with minimal inherent connection to automotive brands. BrandFusion employs a *Lifestyle Representation* strategy, integrating BMW through a detailed miniature car model displayed on a cluttered shelf. The generated video captures the musician's creative space with artistic decorations, music equipment, and personal items, among which the BMW model appears as a subtle reflection of taste and aspirations. The brand presence is understated yet identifiable, avoiding disruption to the scene's artistic focus while successfully incorporating the brand as a natural element of the room's decor.

### F.1.2. IKEA Brand Integration Across Match Levels

Figure 12 showcases three IKEA integration scenarios demonstrating the framework's versatility with home furnishing brands. These cases highlight different placement strategies ranging from functional product usage to ambient brand presence:

**High Match Scenario: Home Organization.** For the prompt "a young woman organizing clothes in a spacious

Table 8. Cost-performance analysis across LLM backbones. Relative inference cost is normalized to GPT-5 as baseline (1.0×). Naturalness Score per dollar measures integration quality efficiency.

| LLM Backbone | Relative Cost | Naturalness Score | NS per $ (Efficiency) |
|---|---|---|---|
| GPT-4o-mini | 0.125× | 4.52 | 36.16 |
| GPT-5 (Default) | 1.00× | 4.70 | 4.70 |
| Gemini-2.5-Pro | 1.00× | 4.91 | 4.91 |

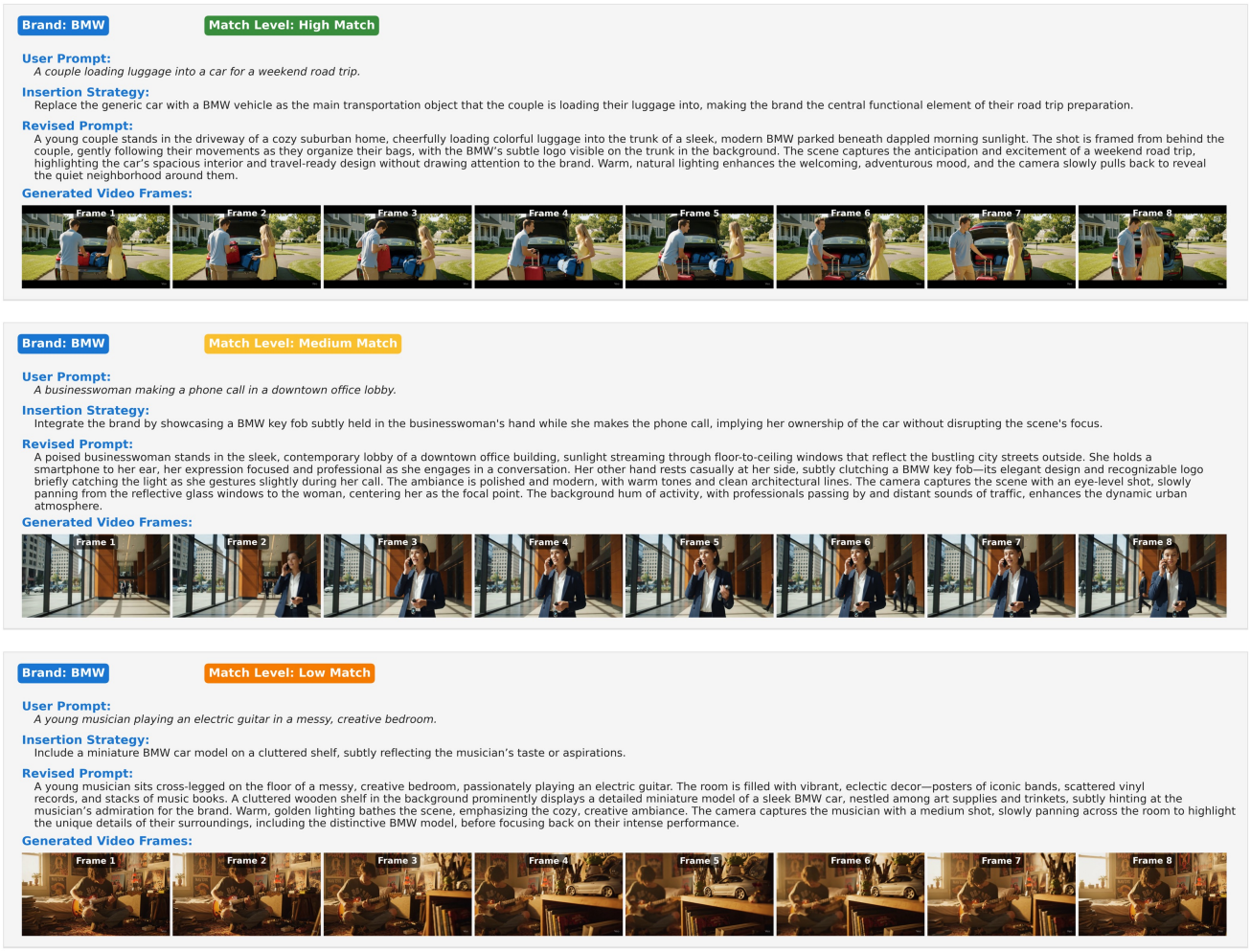


Figure 11. **BMW brand integration examples across different match levels.** (Top) High Match: BMW vehicle serves as the central functional element for a road trip scenario. (Middle) Medium Match: BMW key fob subtly indicates brand ownership in an office setting. (Bottom) Low Match: Miniature BMW model on shelf reflects the musician's taste and aspirations in a creative bedroom scene.

wardrobe at home," BrandFusion identifies strong semantic compatibility and employs a *Functional Product Integration* strategy. The generated video features the woman using distinctive IKEA storage boxes with the classic yellow and blue logo to organize her wardrobe. The boxes are shown being filled with folded sweaters and placed on shelves, naturally showcasing their functionality. Natural lighting through a nearby window illuminates the IKEA branding, achieving clear visibility while maintaining focus on the organizational activity. This exemplifies ideal brand-context alignment where the product serves its intended purpose within the scene.

**Medium Match Scenario: Birthday Celebration.** The prompt "children playing board games on a colorful rug during a birthday party" presents moderate integration dif-

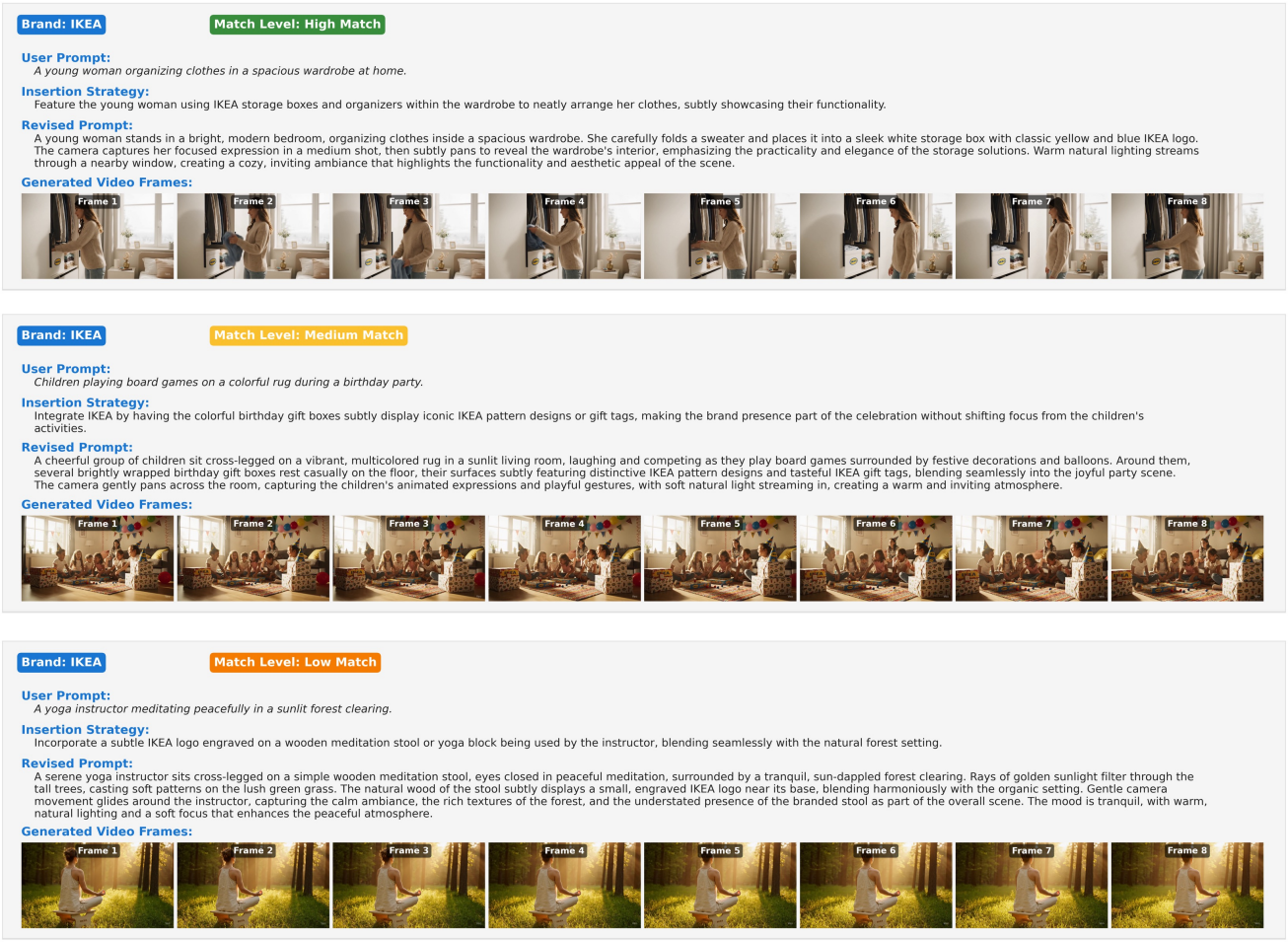


Figure 12. **IKEA brand integration examples across different match levels.** (Top) High Match: IKEA storage solutions serve as functional organizational tools in a home wardrobe scene. (Middle) Medium Match: IKEA patterns appear on birthday party gift boxes, integrating the brand into a celebratory context. (Bottom) Low Match: IKEA logo on a wooden meditation stool blends harmoniously with a natural forest setting.

ficulty as it lacks explicit furniture context. BrandFusion adopts an *Environmental Integration* strategy, incorporating IKEA through decorative elements. The generated video shows a vibrant birthday party scene with festive decorations and balloons, where several wrapped gift boxes displaying iconic IKEA patterns rest casually near the children. The IKEA branding appears on the gift wrap's distinctive designs and tags, making the brand present as part of the celebration's aesthetic without disrupting focus on the children's activities. This demonstrates creative contextual placement where brand elements enhance scene authenticity.

**Low Match Scenario: Forest Meditation.** The most challenging scenario involves "a yoga instructor meditating peacefully in a sunlit forest clearing"—far removed from typical home furnishing contexts. BrandFusion employs a *Natural Object Integration* strategy, incorporating a wooden meditation stool with a subtle IKEA logo engraved near its base. The generated video captures the serene forest environment with sunlight filtering through trees, where the natural wood of the stool blends harmoniously with the organic setting. The IKEA logo is understated yet identifiable upon closer inspection, integrated as part of the instructor's mindfulness equipment. This case demonstrates sophisticated integration where the brand appears as a natural component of the scene rather than an intrusive commercial element, maintaining the tranquil atmosphere while achieving brand presence.

These detailed examples complement our quantitative evaluation by illustrating the qualitative sophistication of BrandFusion's multi-agent collaboration, showcasing how strategic planning, iterative refinement, and context-aware

reasoning converge to achieve seamless brand integration across varying difficulty levels.

## G. Limitations and Ethical Considerations

While BrandFusion demonstrates effective brand integration capabilities, we acknowledge several important limitations and ethical considerations that warrant careful attention in real-world deployment.

### G.1. Technical Limitations

**Dependency on T2V Model Capabilities.** BrandFusion's integration quality is fundamentally constrained by the underlying T2V model's generation capabilities. When T2V models struggle with specific scene types (e.g., complex physical interactions, fine-grained object details, or rapid motion), brand integration quality correspondingly degrades. Our framework cannot compensate for inherent T2V model limitations in rendering brand elements with high fidelity or maintaining temporal consistency across video frames.

**Challenging Integration Scenarios.** Certain prompt-brand combinations remain fundamentally difficult despite our multi-agent approach. Extremely Low Match scenarios—where semantic compatibility between user intent and brand context is minimal—sometimes result in forced or unnatural integrations that fail to satisfy both semantic preservation and brand visibility requirements. Additionally, prompts involving abstract concepts, purely natural environments without human presence, or historical settings often provide limited natural placement opportunities.

**Multi-Brand Integration Complexity.** The current framework focuses on single-brand integration per video generation request. Extending to simultaneous multi-brand scenarios introduces substantial complexity: brands may compete for visual attention, strategic planning must balance multiple advertiser interests, and evaluation criteria become multidimensional. While conceptually feasible, multi-brand integration requires additional framework development to ensure fair treatment and natural coexistence of multiple commercial elements.

**Cultural and Contextual Sensitivity.** BrandFusion's integration strategies are developed primarily on Western commercial contexts and may not generalize optimally to diverse cultural settings where brand perception, advertising norms, and visual aesthetics differ significantly. Brand placement approaches considered natural in one cultural context might appear inappropriate or ineffective in another. Adapting the framework to respect cultural nuances requires expanding the Brand Knowledge Base with region-specific integration guidelines and cultural sensitivity considerations.

### G.2. Ethical Implications

**User Consent and Transparency.** The most critical ethical consideration concerns user awareness and consent regarding brand integration. Users submitting prompts for video generation have a right to know whether and how commercial brands will be incorporated into their content. We strongly advocate for transparent disclosure mechanisms where users are explicitly informed that generated videos may contain brand placements, with options to opt out or select preferred brands. The ecosystem model presented in Figure 2 assumes user awareness, but real-world implementations must ensure this assumption holds through clear interface design and user agreement protocols.

**Manipulation and Authenticity Concerns.** Seamless brand integration, by design, aims to make advertisements appear natural and unobtrusive. This raises concerns about subliminal influence and the blurring of boundaries between user-generated creative content and commercial messaging. When brand elements are integrated so naturally that users cannot distinguish them from organic scene components, there is potential for manipulative advertising that circumvents critical engagement. Service providers must balance commercial interests with ethical obligations to preserve user autonomy and prevent deceptive practices.

**Impact on Creative Expression.** While BrandFusion prioritizes semantic fidelity to user prompts, brand integration inevitably constrains creative freedom to some degree. Users may feel that commercial elements compromise their artistic vision or dilute their intended message. This tension between monetization and creative autonomy is inherent to ad-supported content models. We recommend providing users with control mechanisms—such as brand category preferences, integration intensity settings, or premium ad-free tiers—to mitigate this concern and respect individual creative priorities.

**Equity and Access Considerations.** Ad-supported T2V services enabled by frameworks like BrandFusion could democratize access to expensive video generation technologies by offsetting computational costs. However, this model may also create inequities where users unable to afford premium ad-free options receive lower-quality experiences with intrusive brand placements. Service providers should carefully balance revenue generation with equitable access, potentially offering free tiers with reasonable integration constraints rather than exploitative advertising loads.

**Data Privacy and Brand Profiling.** To optimize brand selection and integration strategies, systems may collect user preference data, prompt histories, and interaction patterns. This data collection raises privacy concerns regarding user profiling for targeted advertising. Service providers must implement robust data protection measures, clearly communicate data usage policies, and provide users with data control and deletion rights in compliance with privacy regulations (e.g., GDPR, CCPA).

### G.3. Potential Misuse and Safeguards

**Unauthorized Brand Usage.** BrandFusion could potentially be misused to generate videos containing brands without proper authorization or licensing agreements. Malicious actors might integrate competitor brands to create false associations, generate counterfeit content, or produce misleading advertisements. To prevent unauthorized usage, we recommend implementing strict brand verification mechanisms where only registered brands with verified ownership can be integrated. Service providers should maintain comprehensive Brand Knowledge Bases with authenticated brand profiles and legal authorization records.

**Inappropriate Content Associations.** Without proper safeguards, brands could be integrated into inappropriate, offensive, or harmful content that damages brand reputation or violates advertising standards. For example, alcohol brands appearing in content involving minors, or luxury brands associated with illegal activities. Our framework includes a "Prohibited Scenarios" field in each brand profile (Section B.1) to specify contextual restrictions. We strongly recommend expanding this mechanism with automated content moderation systems that detect potentially inappropriate prompt-brand combinations and reject integration requests that violate ethical guidelines or legal restrictions.

**Children's Content Protections.** Brand integration in content targeting or involving children raises heightened ethical concerns and regulatory compliance issues (e.g., COPPA in the United States, GDPR's child protection provisions in Europe). Children may be particularly susceptible to integrated advertising and less capable of recognizing commercial intent. We advocate for strict policies prohibiting brand integration in content explicitly identified as child-directed, implementing age verification mechanisms, and applying enhanced scrutiny to any integration involving minors or child-related contexts.

**Misinformation and Deceptive Content.** The framework could potentially be misused to create deceptive videos where brands are associated with false claims, fabricated testimonials, or misleading contexts. For instance, generating videos that falsely suggest brand endorsements by public figures or misrepresent product capabilities. Safeguards should include content verification systems, explicit disclaimers indicating AI-generated content, and prohibition of integration in news-style or documentary formats where authenticity expectations are high.

**Recommended Safeguard Framework.** Based on these considerations, we recommend service providers implementing BrandFusion adopt a comprehensive safeguard framework:

- **Transparent Disclosure**: Clear visual indicators (watermarks, labels) identifying AI-generated content with brand integration
- **User Consent Mechanisms**: Explicit opt-in requirements and granular control over brand integration preferences
- **Content Moderation**: Automated and human review systems to detect policy violations, inappropriate associations, and harmful content
- **Brand Authorization Verification**: Strict authentication protocols ensuring only legitimately registered brands are integrated
- **Regulatory Compliance**: Adherence to advertising standards, consumer protection laws, and platform-specific policies
- **Appeal and Redress Mechanisms**: Processes for users and brands to report misuse and request content removal
- **Continuous Monitoring**: Regular audits of generated content and integration patterns to identify emerging misuse trends

**Research Community Responsibility.** As researchers introducing brand integration capabilities for T2V generation, we acknowledge responsibility for anticipating and mitigating potential harms. We encourage the research community to engage in ongoing dialogue about ethical AI deployment in advertising contexts, develop shared best practices, and collaborate with policymakers to establish appropriate regulatory frameworks. The goal should be creating sustainable T2V monetization pathways that respect user rights, protect vulnerable populations, and maintain public trust in AI-generated media.

In summary, while BrandFusion offers technical solutions for seamless brand integration, successful real-world deployment requires careful attention to technical limitations, ethical implications, and potential misuse scenarios. Service providers must implement comprehensive safeguards, maintain transparency with users, and prioritize ethical considerations alongside commercial objectives. Only through responsible deployment can brand integration contribute positively to sustainable T2V service ecosystems.