

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

Zihao Zhu¹ Ruotong Wang¹ Siwei Lyu³ Min Zhang^{4,2} Baoyuan Wu^{1,2*}

¹School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China

²Shenzhen Loop Area Institute

³State University of New York at Buffalo

⁴School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

{zihaozhu, ruotongwang1}@link.cuhk.edu.cn;

siweilyu@buffalo.edu; zhangmin2021@hit.edu.cn; wubaoyuan@cuhk.edu.cn

Website: <https://zihao-ai.github.io/brandfusion>



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 frame-

work 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

* Corresponding Author.

1. Introduction

The rapid advancement of text-to-video (T2V) generation models, such as Veo [8], Sora [28], and Kling [21], has 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 DreamBooth [32] and Textual Inversion [5] enable models to synthesize novel entities through lightweight fine-tuning. Recent work has attempted brand embedding 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.

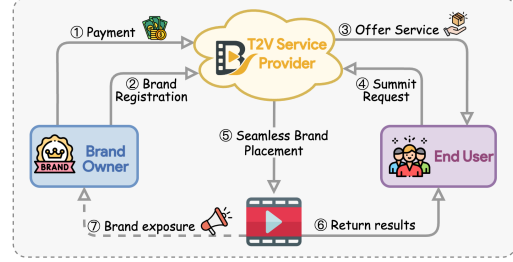


Figure 2. Ecosystem of brand integration in T2V generation.

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.

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 suc-

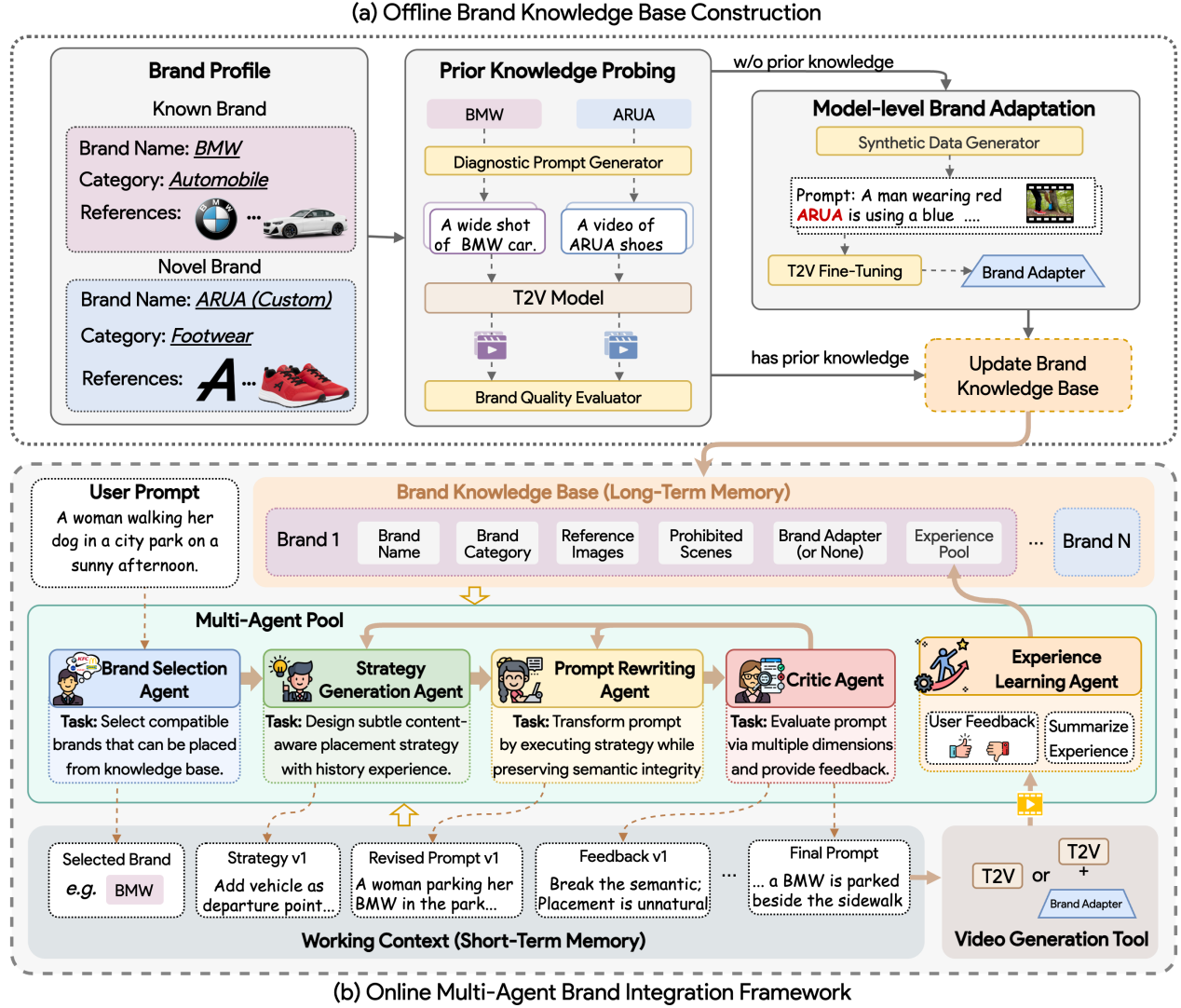


Figure 3. Overview of the BrandFusion framework: (a) Offline phase builds brand knowledge through probing and adaptation. (b) Online phase employs five collaborative agents for semantic-preserving brand integration with continuous learning.

successful integration cases, serving as the foundation for subsequent online integration.

Prior Knowledge Probing. To determine whether the T2V 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 in-

ject 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 $\mathcal{A}_B$. 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 re-

sults 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 prompt-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].

Baseline 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-intrusiveness. 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 1×10^{-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 compro-

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

mise 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.

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-

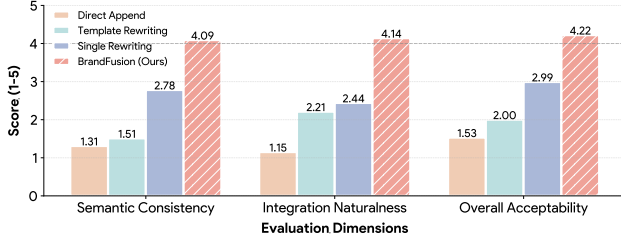


Figure 4. Human evaluation results on semantic fidelity, integration naturalness, and overall acceptability.

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 model¹. 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

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

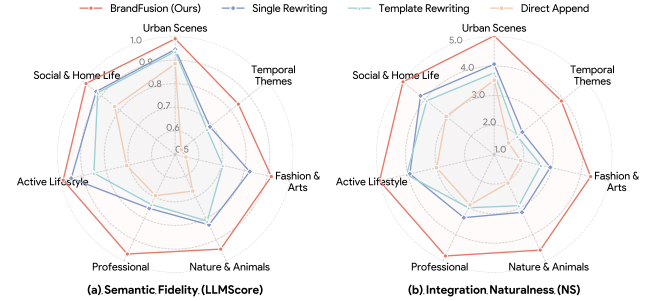


Figure 5. Comparison across different prompt scene categories.

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.

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 10 groups to compute average overall acceptability within

¹This study was approved by our institution’s IRB.



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.

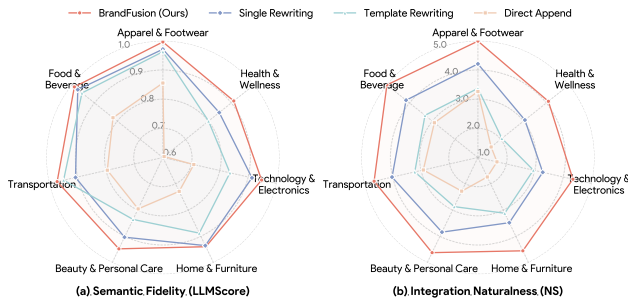


Figure 7. Comparison across seven brand categories.

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

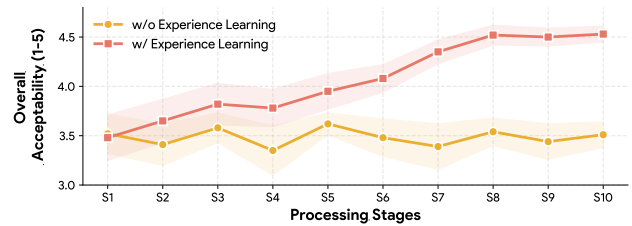


Figure 8. Experience learning effectiveness over sequential stages.

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 categories. 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.

Acknowledgments

Baoyuan Wu is supported by Guangdong Basic and Applied Basic Research Foundation (No. 2024B1515020095), Guangdong Provincial Program (No. 2023TQ07A352), Shenzhen Science and Technology Program (No. RCYX20210609103057050 and JCYJ20240813113608011), and Longgang District Key Laboratory of Intelligent Digital Economy Security.

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. 3
- [6] Jiaojia 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. 2, 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. 2, 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 Serkan Ö Arik. 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. 2, 5
- [29] YP Pragathi, Swaroop S. Bharadwaj, Preetham Kolli, S Sanjith, Kavitha Sooda, and S Sunayana. Adgen: A multi-modal 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. 3
- [33] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyan 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, Jingen 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, Wenten Wang, Wenting Shen, Wenyan 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. Modelscape 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, Yanan 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, Hao-long 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](#)