PROAGENT: HIGH-FIDELITY PRODUCT VIDEO GENERATION VIA A MULTI-AGENT FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent text-to-video models have made advertising video generation attractive as a way to cut production time and cost, yet direct prompting often yields videos that drift from product identity, contain hallucinations, distort printed text, and fail to convey a clear commercial expression. We present **ProAgent**, the first to address high-fidelity product video generation—a training-free multi-agent framework that turns a user prompt identifying a specific product into an advertisement-ready video end-to-end. ProAgent grounds this intent in retrieved product evidence, plans a shot-level blueprint and rewrites it into a generation prompt under explicit product constraints, and iteratively improves the result through a closed-loop critique-and-recovery process. To evaluate the quality of generated advertising videos, we further introduce **PV-Bench** (Product Video Benchmark), a multi-dimensional benchmark scoring videos across six product-centric dimensions. Experiments across 12 product categories spanning major e-commerce industries and multiple state-of-the-art video generators show that ProAgent consistently outperforms baselines on both PV-Bench and general metrics for video quality and semantic fidelity.

1 INTRODUCTION

Product advertising videos have become a primary medium for online product promotion, yet producing them relies on professional teams for commercial photography, scene construction, and post-production editing, making every clip both time-consuming and expensive. The rapid progress of text-to-video (T2V) generation (OpenAI, 2025; DeepMind, 2025; Seedance et al., 2026; Kong et al., 2024; MiniMax, 2026) has made AI-driven video creation widely accessible, yet its outputs still fall short of commercial advertising standards. A few efforts connect video generation to commercial settings: BrandFusion (Zhu et al., 2026) embeds brand elements into generic scenes, and AdsQA (Long et al., 2025b) studies advertisement video understanding—yet neither targets faithful product-centric advertising *generation*. In parallel, multi-agent video frameworks (Huang et al., 2026; Wu et al., 2025; Hu et al., 2024; Huang et al., 2024a) improve narrative coherence and compositional alignment, but operate on *generic* content without enforcing product fidelity. Together, these leave faithful product-centric advertising video generation unaddressed. We therefore introduce, for the first time, the task of *high-fidelity product video generation*: automatically producing product-faithful advertising videos from a user's intent.

Unlike generic video generation, where any visually plausible output is acceptable, a product advertising video must be anchored to a real, existing product: any deviation in color, shape, material, logo, or on-product text may mislead consumers and damage brand trust. Generating such videos presents challenges that extend far beyond prompt engineering: (1) *Product grounding*: a user prompt conveys only coarse intent, so the system must first locate the correct product information and select the most representative views from a pool of candidates. (2) *Product fidelity*: the video must preserve the color, shape, material, and multi-view consistency of these references. (3) *Hallucination robustness* and *text accuracy*: no object, action, or accessory absent from the references may appear, and product text and logos must remain clear and unmodified. (4) *Commercial expression*: the video must present the product in an appealing, purchase-motivating way. As shown in Figure 1, standard generation pipelines still fail to meet these demands: (a) prompting directly yields the wrong product, and (b) even conditioning on a retrieved reference image leaves appearance drift and fidelity failures.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Illustration of high-fidelity product video generation using a bracelet. **(a)** Direct Prompting generates the wrong product. **(b)** Adding a retrieved reference image improves product identity but still introduces visual inconsistencies. **(c)** ProAgent produces a faithful, advertisement-ready video.

To address these challenges, we propose **ProAgent**, a training-free multi-agent framework for high-fidelity product video generation. Our key insight is that faithful product video synthesis is better served by closed-loop generation with iterative review and repair (Madaan et al., 2023; Shinn et al., 2023) than by one-shot sampling, and that its two prerequisites—grounded product knowledge and actionable visual diagnosis—are naturally provided by a set of specialized, collaborating agents. As illustrated in Figure 2, ProAgent operates through three cooperative stages. *(i) Context grounding.* Instead of handing the raw prompt to the generator, ProAgent analyzes the user intent to extract the target product information and the desired video characteristics, then retrieves product evidence from external image search and web search. *(ii) Planning and generation.* A Planner Agent decomposes the advertising intent into a structured shot blueprint that specifies the shot type, selling point, anchored references, duration, and scene action; a Prompt Rewriter Agent then converts the blueprint into a generation prompt under skill-guided constraints. *(iii) Review and recovery.* Because a single sample cannot guarantee fidelity, a Critic Agent diagnoses the generated video along multiple product-centric dimensions, returning a defect report rather than a scalar score; a Recovery Agent selects a targeted action according to the report—edit the prompt or revise the plan—while an Experience Learning Agent distills reusable experience from the iteration trajectories. All modules run at inference time and update no parameters, so ProAgent applies to any black-box generator.

To enable comprehensive evaluation of this task, we introduce **PV-Bench**, a benchmark of 120 real e-commerce products with an MLLM-as-Judge evaluation protocol. It scores generated videos across six product-centric dimensions and aggregates the resulting scores into Pass Rate (PR), Checklist Accuracy (CA), and the weighted PV-Score. Our main contributions are three-fold:

- We introduce high-fidelity product video generation as a novel task, accompanied by PV-Bench, an MLLM-as-Judge evaluation protocol that scores product videos along six product-centric dimensions.
- We propose ProAgent, the first training-free multi-agent framework for this task, featuring three cooperative stages: context grounding with product-aware retrieval, planning and generation with structured shot planning and constrained prompt rewriting, and review and recovery with multi-dimensional critique, targeted repair, and experience reuse.
- We conduct extensive experiments on real e-commerce products spanning the major verticals of online retail across multiple T2V models, achieving state-of-the-art performance on both our product-centric PV-Bench and general video evaluation metrics.

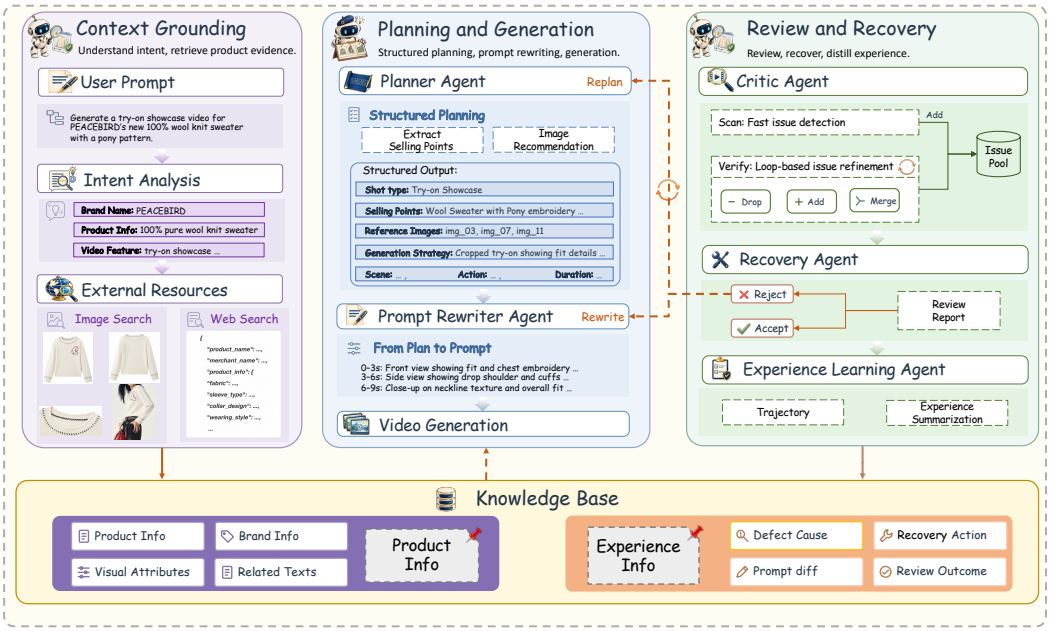

Figure 2: Overview of ProAgent. (a) Context Grounding: parses user intent and retrieves product evidence. (b) Planning and Generation: produces a structured shot blueprint, rewrites it into a constrained prompt, and generates the video. (c) Review and Recovery: reviews the result, repairs failures, and distills reusable experience.

2 Related Work

2.1 Video Generation

Recent diffusion-based video generation models have substantially improved natural video generation in terms of quality, length, and controllability (OpenAI, 2025; DeepMind, 2025; Seedance et al., 2026; HappyHorse, 2026; MiniMax, 2026; Kong et al., 2024; Wan et al., 2025; Yang et al., 2024). However, these end-to-end models still struggle with long videos and multi-shot composition (Liu et al., 2026), multi-image conditioning, and, crucially, fine-grained subject consistency. Even when reference images are supplied, subject-driven video generation still struggles to preserve the referenced subject's identity and appearance consistently across frames (Yuan et al., 2025; Liu et al., 2025). These limitations motivate a structured, reference-aware generation pipeline—in the spirit of retrieval-augmented generation (Chen et al., 2022)—rather than a single end-to-end inference call.

2.2 Multi-Agent Frameworks for Video Creation

Professional video production is organized as a pipeline of specialized roles—scripting, directing, shot planning, and editing—and recent work mirrors this structure with multi-agent collaboration, assigning each role to a dedicated agent to improve narrative coherence and production quality. ViMax (Huang et al., 2026) and MovieAgent (Wu et al., 2025) assign director, screenwriter, and designer roles to generate long-form videos with cross-shot consistency, while Paper2Video (Zhu et al., 2025) and Code2Video (Chen et al., 2025) generate academic and educational presentation videos, and GenMAC (Huang et al., 2024a), StoryAgent (Hu et al., 2024), V-Stylist (Yue et al., 2025), and VISTA (Long et al., 2025a) improve generic content through iterative refinement loops—via redesign, agent reflection, and test-time self-improvement. BrandFusion embeds brand elements into generic scenes through a multi-agent pipeline (Zhu et al., 2026) rather than generating faithful product-centric showcase videos. Unlike these methods, we are the first to introduce the task of product-faithful advertising video generation, and we address it with ProAgent, a training-free multi-agent framework, together with PV-Bench, a multi-dimensional benchmark for evaluating product advertising videos.

3 PROAGENT

3.1 OVERVIEW AND PROBLEM FORMULATION

Our goal is to automatically turn a user intent that names a product into an advertising video, so that the video follows what the user asked for while putting the product itself at the center of the presentation. Handing such an intent straight to a T2V model, however, runs into several obstacles: the product evidence required for generation is scattered and hard to retrieve, and the resulting videos drift from product identity, contain hallucinations, distort printed text, and fail to convey a clear commercial expression. We therefore propose ProAgent, a training-free multi-agent framework for high-fidelity product video generation. As illustrated in Figure 2, it consists of three cooperative stages: (1) *Context Grounding*, which parses the user intent and consolidates retrieved product evidence into a product knowledge base; (2) *Planning and Generation*, which commits the grounded intent to a shot-level blueprint and materializes it as a generation prompt under explicit product constraints; and (3) *Review and Recovery*, which audits each generated video through a coarse-to-fine scan-and-verify process and applies a targeted corrective action before regenerating. An Experience Learning Agent closes the loop: it distills its own iteration trajectories—both effective fixes and recurring pitfalls—into reusable prompt patterns that the Rewriter retrieves to strengthen future generations.

3.2 CONTEXT GROUNDING

Intent Analysis. A user starts by stating a creative idea in natural language. Such an input I arrives in diverse narrative forms, and the information about the product is often scattered across the description. To settle the product subject and the user intent, ProAgent first converts I into structured information,

$$R = \mathrm{IntentAnalysis}(I), \tag{1}$$

where $\mathrm{IntentAnalysis}(\cdot)$ is an LLM-based entity recognition step that extracts structured fields from I, such as the target product name, the brand name, and the desired video characteristics, which guide the subsequent production.

Product Evidence Retrieval. The structured intent R captures whatever the user has stated, yet producing a faithful video calls for richer visual and textual evidence about the product than a prompt usually carries. ProAgent therefore grounds R in evidence collected from the open web. Taking the extracted product and brand names as queries, it performs two complementary retrievals,

$$E_t = \mathrm{WebSearch}(R), \qquad E_v = \mathrm{ImageSearch}(R), \tag{2}$$

where $\mathrm{WebSearch}(\cdot)$ collects textual evidence such as specifications, selling points, and brand descriptions, and summarizes it into concise product facts, while $\mathrm{ImageSearch}(\cdot)$ casts a wide net over product images across diverse views and usage scenes, each captioned by an MLLM into structured visual attributes. These retained images—far more than any single video will use—are tagged with the shot types they can support, and together with the textual facts constitute the product knowledge base $\mathcal{K} = (E_t, E_v)$ on which all downstream agents operate.

3.3 PLANNING AND GENERATION

Structured Shot Planning. The Planner Agent turns the grounded context into a single shot blueprint that fixes what the video should show,

$$\mathcal{B} = \mathrm{ShotPlanning}(R, \mathcal{K}). \tag{3}$$

Guided by the user intent R, it first settles the shot type—the concrete form the video takes—and retrieves, among the images tagged with that shot type, the most relevant ones as the anchoring references. It then composes the selling point, a full sentence that tells the downstream rewriter what the video should convey, grounded in the product evidence $\mathcal{K}$. Beyond the shot type, the selling point, and the anchoring references, the blueprint $\mathcal{B}$ also fixes a generation strategy, the duration, the scene and primary action, and whether a specific human identity is required, serving as the universal anchor for the downstream modules.

Skill-Guided Prompt Rewriting. The Prompt Rewriter Agent decides only *how to express* the plan, translating the blueprint into a generation-ready prompt while leaving the creative decisions to

the planner. Rewriting runs in two modes. On the first attempt, it expands the blueprint into a rich, self-contained prompt that spells out every element the video needs, guided by an experience pool $\mathcal{E}$ distilled from past generations,

$$p_0 = \mathrm{Rewrite}(\mathcal{B}, \mathcal{E}). \tag{4}$$

On a later attempt k, rather than starting over, it revises the previous prompt against the diagnostic feedback accumulated so far, correcting the criticized aspects,

$$p_k = \mathrm{Rewrite}(\mathcal{B}, \mathcal{H}_{k-1}, \mathcal{E}), \tag{5}$$

where $\mathcal{H}_{k-1} = \{(p_j, c_j)\}_{j<k}$ collects the earlier prompts p_j and their critique reports c_j. In both modes, the agent draws on reusable prompt-writing skills and finishes with a checklist-based self-check, revising the draft before output. The finalized prompt is then submitted to a black-box text-to-video generator $\mathcal{G}$ to synthesize the candidate video for review,

$$V_k = \mathcal{G}(p_k). \tag{6}$$

3.4 Review and Recovery

Scan-and-Verify Critic. For our product-advertising video review task, a single-pass VLM reviewer trades off between over-flagging trivial blemishes and missing core defects. The Critic therefore follows a coarse-to-fine, scan-then-verify design (Figure 2). In the *Scan* stage, a team of specialized sub-critics—several checks over a fixed set of predefined issues, plus one open-ended check for anything beyond them—quickly scans sparsely sampled frames of the video against $\mathcal{B}$ and $\mathcal{K}$, recalling candidate issues into an initial pool $\mathcal{I}_0$. The *Verify* stage then re-examines each pooled issue on more densely sampled frames with a ReAct loop (Yao et al., 2022)—dropping unfounded false positives, adding overlooked defects, and merging duplicates that share a root cause—and iterates until the issue pool stabilizes, producing a refined review report c for the downstream Recovery Agent:

$$\mathcal{I}_0 = \mathrm{Scan}(V_k, \mathcal{B}, \mathcal{K}), \qquad c = \mathrm{Verify}(\mathcal{I}_0, V_k, \mathcal{B}, \mathcal{K}). \tag{7}$$

Diagnosis-Driven Recovery. When the Critic rejects a video, the Recovery Agent adaptively selects a recovery action—based on the review report, the history of attempts, and the remaining budget: it either (i) *edits the prompt*, letting the Prompt Rewriter Agent fix only the flagged defects, or (ii) *revises the plan*, returning to the Planner Agent to swap the primary reference image or replan the shot strategy. The chosen action is recorded in a revision history and fed back to the corresponding agent in the next iteration. This process continues until a stopping criterion is met or the maximum number of attempts is reached.

Experience Learning. ProAgent turns each shot's own iteration trajectory into reusable experience: the generate-and-review attempts expose which prompt patterns trigger defects and, when a repair passes, which revision removes them. The Experience Learning Agent distills these signals into a product-agnostic entry that pairs the triggering context with one of two lessons—(1) a verified fix, when a repair succeeds, or (2) a pattern to avoid, when none pass—and deposits it in the experience pool $\mathcal{E}$ that the Prompt Rewriter Agent retrieves in later generations, so even the first prompt p_0 avoids pitfalls seen before. Through this closed-loop learning, ProAgent improves continuously across tasks without updating the video generator, keeping the framework training-free.

4 Experiments

4.1 Evaluation Setup

Product Video Bench. To evaluate ProAgent's ability to generate faithful product advertising videos, we construct PV-Bench (Product Video Benchmark) from 120 real e-commerce products, 10 from each of 12 industry categories spanning the major verticals of online retail (e.g., home appliances, consumer electronics, apparel and fashion, and food and fresh produce). Each video is scored by an MLLM-as-Judge checklist (Chen et al., 2024), following the checklist-based paradigm of Qwen-Image-Agent's IA-Bench (Zhang et al., 2026), along six product-centric dimensions: Product Fidelity ($\mathcal{F}$), Hallucination Robustness ($\mathcal{H}$), Text Accuracy ($\mathcal{T}$), Physics Realism ($\mathcal{P}$), Commercial Expression ($\mathcal{C}$), and Aesthetics ($\mathcal{A}$). Each dimension contains 4–5 binary items verified by an MLLM, and we report Pass Rate (PR), the fraction passing all items, and Checklist Accuracy (CA),

Table 1: Main results on PV-Bench for three video generators. Checklist Accuracy (CA) and Pass Rate (PR) are reported in separate column blocks for the six product-centric dimensions. PF: Product Fidelity, HR: Hallucination Robustness, TA: Text Accuracy, PHY: Physics Realism, CE: Commercial Expression, AES: Aesthetics. Best per generator in **bold**.

Model	Method	Checklist Accuracy (%)						Pass Rate (%)						PV-Score
		PF	HR	TA	PHY	CE	AES	PF	HR	TA	PHY	CE	AES	
MiniMax-H3	DP	81.8	78.3	75.0	**99.0**	68.1	92.7	51.7	51.7	60.8	96.7	45.0	76.7	81.0
	SR	88.3	90.0	87.1	98.5	71.5	87.3	69.2	77.5	80.0	94.2	44.2	60.8	87.7
	Ours	**90.7**	**93.3**	**89.4**	**99.0**	**81.9**	**93.8**	**73.3**	**82.5**	**83.3**	**97.5**	**67.5**	**85.0**	**91.3**
HappyHorse	DP	79.2	82.9	68.5	98.7	76.5	92.3	43.3	60.0	59.2	96.7	56.7	79.2	81.0
	SR	78.8	87.5	64.8	97.2	68.5	89.4	44.2	68.3	53.3	89.2	47.5	69.2	80.1
	Ours	**84.5**	**88.8**	**73.5**	**99.7**	**77.5**	**93.1**	**48.3**	**70.8**	**60.8**	**98.3**	**61.7**	**81.7**	**85.1**
Seedance	DP	87.5	84.8	73.7	**99.5**	84.6	96.4	60.7	64.1	61.5	97.4	59.0	94.0	85.9
	SR	88.3	88.6	79.7	97.5	71.4	85.6	66.1	74.6	71.2	90.7	40.7	70.3	85.6
	Ours	**92.0**	**93.3**	**82.5**	**99.5**	**87.3**	**97.1**	**73.3**	**78.3**	**74.2**	**97.5**	**72.5**	**95.0**	**91.2**

the average fraction of satisfied items. To prevent self-evaluation bias, all methods are scored by a separate judge model, Gemini-3.1-pro (distinct from the agent backbone), using a fixed checklist. We further aggregate the six CA scores into a single Product Video Score,

$$\text{PV-Score} = 0.25\,\mathcal{F} + 0.25\,\mathcal{H} + 0.20\,\mathcal{T} + 0.10\,\mathcal{P} + 0.10\,\mathcal{C} + 0.10\,\mathcal{A}, \tag{8}$$

where the weights reflect the relative importance of each dimension for downstream commercial usability.

Baseline Methods and Model Selection. We compare ProAgent against two baselines that reflect current practice: (1) Direct Prompting (DP) directly instructs the generator to produce a product advertising video from the raw user prompt; (2) Single Rewriting (SR) employs an LLM for one-pass rewriting of the product description into a single generation prompt, without any critique or recovery. Notably, for a fair comparison we also supply DP and SR with the top-1 most relevant retrieved reference image: without it, they feed text directly to the video generator, which lacks world knowledge of the specific product and renders the wrong product. Since ProAgent also relies on reference images to anchor the product, this equal grounding lets the comparison isolate the gains of its multi-agent planning, critique, and repair over the single-pass baselines. To further assess generalization across model families, we instantiate all methods on three state-of-the-art video generators: MiniMax-H3 (MiniMax, 2026), an open-source video generation model, and two closed-source commercial models, HappyHorse (HappyHorse, 2026) and Seedance 2.0 (Seedance et al., 2026). For brevity we refer to Seedance 2.0 as Seedance in the tables below.

Evaluation Metrics. We evaluate generated videos from three complementary perspectives. (a) **Product advertising video quality.** We report PR, CA, and the aggregated PV-Score from our six-dimensional MLLM-as-Judge checklist, which directly measures product-level fidelity and commercial usability. (b) **General video quality.** We adopt VBench-Quality (Huang et al., 2024b), computed as the arithmetic mean of the seven VBench Technical Quality dimensions (subject consistency, background consistency, temporal flickering, motion smoothness, dynamic degree, aesthetic quality, and imaging quality); and VisionReward (Xu et al., 2024), a VLM-based reward model that assesses text alignment, composition, detail fidelity, and physical realism. (c) **Semantic fidelity.** We use VQAScore to measure semantic alignment via prompt-derived question answering on the generated video, and DINO Sim. (Oquab et al., 2023) to measure appearance consistency between the reference product images and generated video frames.

4.2 Main Results

Results on PV-Bench. As reported in Table 1, ProAgent obtains the best PV-Score on every generator, improving over Direct Prompting by up to $+10.3$ points. The gains concentrate on product fidelity and hallucination robustness, the two dimensions that define the task, indicating that diagnosis-driven, verified correction fixes precisely the failures that a single open-loop rewrite cannot catch. Under the strict Pass Rate criterion, where a single violated checklist item disqualifies

Table 2: General video evaluation metrics. VBench: VBench-Quality; VisRwd: VisionReward; VQA: VQAScore; DINO Sim.: DINOv2 feature similarity to the reference images. VBench, VQA, and DINO Sim. are reported on a 0–100 scale, whereas VisRwd is reported on its original scale.

Model	Method	VBench↑	VisRwd↑	VQA↑	DINO Sim.↑
MiniMax-H3	DP	70.02	0.154	83.95	48.14
	SR	71.05	0.181	89.27	64.01
	Ours	**71.34**	**0.184**	**90.15**	**64.83**
HappyHorse	DP	70.60	0.147	87.46	59.68
	SR	70.86	0.169	91.83	59.80
	Ours	**70.93**	**0.178**	**92.01**	**63.21**
Seedance	DP	68.80	0.160	90.50	53.71
	SR	69.44	0.178	90.91	67.44
	Ours	**69.89**	**0.187**	**92.14**	**70.65**

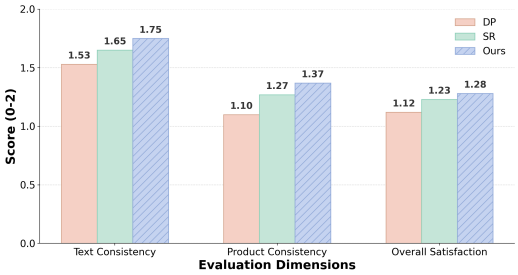

Figure 3: Human evaluation on text consistency, product consistency, and overall satisfaction, scored on a 0–2 scale and aggregated by majority vote across five participants over 60 products. Higher is better.

Table 3: Component ablation on the PV-Bench subset using MiniMax-H3. Each ablated variant removes one ProAgent module. All values in %.

		Checklist Accuracy (%)						
Variant	VBench	PF	HR	TA	PHY	CE	AES	PV-Score
ProAgent (Full)	**71.39**	**93.67**	90.83	90.42	**100.00**	**83.75**	**94.58**	**92.04**
w/o Planner	70.69	87.00	92.08	91.25	97.67	73.33	87.92	88.91
w/o Critic (single-pass)	71.31	89.33	87.92	85.83	**100.00**	**83.75**	93.75	89.23
w/o Experience Learning	71.08	89.33	**92.92**	**92.08**	99.00	75.00	90.42	90.42

a video, the margins widen further, showing that ProAgent passes more videos in full rather than merely raising per-item scores. In contrast, Single Rewriting is not a reliable improvement over DP: this unverified rewrite trades detail for fidelity—Hallucination Robustness rises on all three generators—but crowds out the budget for cinematography, lowering Aesthetics everywhere and hurting Commercial Expression and Physics Realism on the two stronger models. It thus raises PV-Score only on the weakest generator (MiniMax-H3) and stays flat or negative on HappyHorse and Seedance, as its fidelity gains are canceled by the regressions they cause.

Results on General Metrics. As shown in Table 2, ProAgent ranks first on VBench-Quality and VisionReward for every generator, indicating that the framework as a whole improves generic video quality rather than trading it for product-centric quality. ProAgent also leads on semantic fidelity on every generator: VQAScore shows closer adherence to the generation prompt, and DINO Sim. closer appearance consistency with the reference images. Being feature-based and free of model judgment, DINO Sim. independently corroborates PV-Bench's product fidelity gains.

Human Evaluation. To validate ProAgent from the user perspective, five participants rate anonymized videos from all three methods for 60 products, presented with the reference images in random order. On a three-point scale (0/1/2), they assess *Text Consistency* (the accuracy and legibility of product text), *Product Consistency* (faithful product appearance without hallucinated parts), and *Overall Satisfaction*. As shown in Figure 3, ProAgent scores highest on all three, confirming improved product faithfulness and user satisfaction.

4.3 Ablation Studies

Unless stated otherwise, all ablations use a random half of PV-Bench (60 products), MiniMax-H3 as the generator, and at most $K = 3$ recovery iterations.

Components Ablation. Table 3 isolates the contribution of each module by removing one at a time. Disabling the Planner causes the largest drop in product fidelity and also weakens commercial expression, confirming that structured shot planning and reference-view selection are what ground generation in the real product. Disabling the Critic lowers product fidelity together with hallucination robustness and text accuracy, the failure modes that closed-loop review and recovery are de-

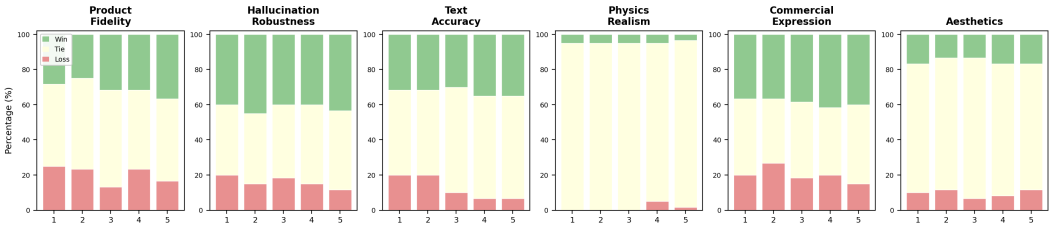

Figure 4: Win/tie/loss rates (%) of ProAgent versus Direct Prompting (DP) across the six PV-Bench dimensions for $K \in \{1, \ldots, 5\}$, based on Checklist Accuracy with MiniMax-H3.

Table 4: Impact of the LLM backbone on the PV-Bench subset with MiniMax-H3 as the video generator. All values in %.

		Checklist Accuracy (%)						
LLM Backbone	VBench	PF	HR	TA	PHY	CE	AES	PV-Score
GPT-5 (default)	**71.39**	**93.67**	**90.83**	**90.42**	**100.00**	**83.75**	**94.58**	**92.04**
Claude-Sonnet-4.5	68.83	73.67	84.17	87.92	99.33	72.50	85.83	82.81
Qwen3.7-Plus	68.15	71.67	81.67	85.00	**100.00**	76.25	90.42	82.00

signed to catch. Disabling Experience Learning also reduces product fidelity while mainly harming commercial expression and aesthetics, suggesting that accumulated experience transfers effective presentation and framing conventions. Individual variants occasionally lead on a single dimension, but only by sacrificing other product-critical objectives, whereas the full model balances all six and attains the best overall PV-Score. VBench remains nearly identical across variants, showing that our gains come from product-level quality rather than generic video quality.

Fine-Grained Results. To better understand the specific dimensions improved by ProAgent, we present a fine-grained win/tie/loss comparison with Direct Prompting across $K \in \{1, \ldots, 5\}$ recovery iterations in Figure 4. ProAgent wins more frequently than it loses across all six PV-Bench dimensions. The most notable gains appear in Product Fidelity, Hallucination Robustness, Text Accuracy, and Commercial Expression, attributable to ProAgent's built-in constraints and iterative refinement process. Finally, ProAgent holds a slight edge on Physics Realism and Aesthetics, as current generators already perform well on both, leaving limited headroom for improvement.

Impact of LLM Backbone. To study the impact of the LLM backbone, we conduct an ablation on the backbone choice. By default, ProAgent uses GPT-5; in the ablation setting, we replace it with Claude-Sonnet-4.5 and Qwen3.7-Plus while keeping all prompts, skills, and hyperparameters fixed. As shown in Table 4, replacing the default backbone causes substantial degradation across most metrics, demonstrating that backbone capability is critical to the overall system. In particular, the largest drops occur in Product Fidelity, indicating that strong planning, instruction following, and visual reasoning are essential for grounding the product constraints and producing accurate critique–recovery decisions. In contrast, Physics Realism remains nearly unchanged, as it is primarily determined by the underlying video generator rather than the agentic pipeline.

4.4 Analysis

Case Studies. Figure 5 shows two product videos generated by ProAgent. For the *PEACEBIRD sweater*, ProAgent preserves the red tone, V-neck cut, and ribbed-knit texture across front, side, and close-up shots, with a model naturally showcasing its fit. For the *MAC cream blush*, the video keeps the brand logo and packaging legible and the shades faithful, progressing from a product display to a color swatch and an on-cheek application that conveys the usage effect. Both examples show that, while faithfully following the user's requirements, ProAgent produces advertisement-ready videos that keep product identity intact—color, shape, texture, and printed logos—free of hallucinated parts or functions and with clear commercial expression.



Figure 5: Case studies of ProAgent-generated product videos, showing faithful product identity and clear, advertisement-ready presentation.

Performance Across Product Categories. To evaluate ProAgent's generalizability across diverse commercial domains, we analyze Checklist Accuracy and Pass Rate on eight representative product categories, using MiniMax-H3. As shown in Figure 6, ProAgent consistently outperforms all baselines on both metrics. ProAgent maintains a clear and stable advantage across all categories, whereas baseline performance varies substantially. These results show that our product-aware multi-agent framework generalizes across diverse product constraints and retail categories.

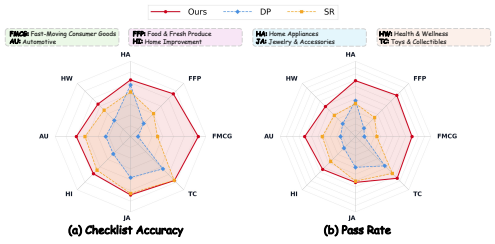

Figure 6: Comparison across different product categories.

5 Conclusion

In this paper, we introduced *high-fidelity product video generation*, a novel task that turns a user prompt specifying a product into an advertisement-ready video, together with **PV-Bench**, an MLLM-as-Judge benchmark scoring real e-commerce products across six advertising-relevant dimensions. To tackle this task, we proposed **ProAgent**, a training-free multi-agent framework for high-fidelity product video generation. Built on shared multimodal product evidence, ProAgent coordinates specialized agents for shot planning, prompt generation, visual review, and targeted recovery; an additional Experience Learning Agent distills reusable knowledge from the closed-loop refinement process. Extensive experiments show that ProAgent consistently outperforms the baselines across state-of-the-art generators, leading on both the product-centric dimensions of PV-Bench and general metrics of video quality and semantic fidelity, improving product fidelity without sacrificing generic quality. Human evaluations further confirm that ProAgent achieves higher user satisfaction. Future work could extend ProAgent to multi-product composition and personalized video generation.

AI USE STATEMENT

In this work, generative AI tools were used solely for language polishing, grammar correction, and improving the overall readability and flow of the text. We have not used generative AI tools for any core research tasks (such as ideation, theoretical derivation, experimental design, code generation, or data analysis), and all other tasks with required disclosure are not applicable to this work.

We have thoroughly reviewed all AI-assisted text: all rephrased sentences and suggestions were manually verified by the authors to ensure that the original scientific meaning and technical claims remain completely unaltered and accurate. The authors take full responsibility for the final content of this manuscript.

ETHICS STATEMENT

The data used in this work are drawn from publicly available real-world product information and are used solely for academic research, without any commercial use. The human evaluation was conducted anonymously with the informed consent of the participants and did not collect any sensitive personal information. We have taken care to avoid potential privacy and intellectual-property risks, comply with the ICLR Code of Ethics, and did not identify any significant ethical concerns.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide in the appendix the necessary details on dataset construction, the evaluation protocol, metric configurations, and the complete prompt templates. The models used and the key hyperparameters are also specified in the main text and the appendix to support reproduction of our results.

REFERENCES

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark, 2024. URL https://arxiv.org/abs/2402.04788.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator, 2022. URL https://arxiv.org/abs/2209.14491.

Yanzhe Chen, Kevin Qinghong Lin, and Mike Zheng Shou. Code2video: A code-centric paradigm for educational video generation, 2025. URL https://arxiv.org/abs/2510.01174.

DeepMind. Veo: A text-to-video generation system. Technical report, DeepMind, May 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

HappyHorse. HappyHorse: AI video generation platform. https://www.happyhorse.cn/, 2026. Accessed: 2026-09-11.

Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. StoryAgent: Customized storytelling video generation via multi-agent collaboration, 2024. URL https://arxiv.org/abs/2411.04925.

Kaiyi Huang, Yukun Huang, Xuefei Ning, Zinan Lin, Yu Wang, and Xihui Liu. GenMAC: Compositional text-to-video generation with multi-agent collaboration, 2024a. URL https://arxiv.org/abs/2412.04440.

Lingxuan Huang, Sizhe He, Hengji Zhou, Liqiang Nie, Lianghao Xia, and Chao Huang. ViMax: Agentic video generation, 2026. URL https://arxiv.org/abs/2606.07649.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b. URL https://arxiv.org/abs/2311.17982.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. URL https://arxiv.org/abs/2412.03603.

Akide Liu, Jinbo Xing, Chaojie Mao, Ye Li, Zeyu Zhang, Yefei He, Weijie Wang, Zihan Wang, Yu Liu, Gholamreza Haffari, and Bohan Zhuang. ReCA: Multi-shot long video extrapolation via recursive context allocation, 2026. URL https://arxiv.org/abs/2605.26525.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment, 2025. URL https://arxiv.org/abs/2502.11079.

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, 2025a. URL https://arxiv.org/abs/2510.15831.

Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, Yang Liu, Jiaheng Ma, and Bowen Zhou. Adsqa: Towards advertisement video understanding, 2025b. URL https://arxiv.org/abs/2509.08621.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

MiniMax. MiniMax-H3. https://github.com/MiniMax-AI/MiniMax-H3, 2026. Accessed: 2026-09-11.

OpenAI. Sora 2 is here. https://openai.com/index/sora-2/, 2025. Accessed: 2026-09-11.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023. URL https://arxiv.org/abs/2304.07193.

Team Seedance et al. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

Team Wan et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning, 2025. URL https://arxiv.org/abs/2503.07314.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation, 2024. URL https://arxiv.org/abs/2412.21059.

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. URL https://arxiv.org/abs/2408.06072.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022. URL https://arxiv.org/abs/2210.03629.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition, 2025. URL https://arxiv.org/abs/2411.17440.

Zhengrong Yue, Shaobin Zhuang, Kunchang Li, Yanbo Ding, and Yali Wang. V-stylist: Video stylization via collaboration and reflection of mllm agents, 2025. URL https://arxiv.org/abs/2503.12077.

Zekai Zhang, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, Xiao Xu, Yan Shu, Yanran Zhang, Yixian Xu, Yuxiang Chen, Zhendong Wang, Zihao Liu, Zikai Zhou, Huishuai Zhang, Dongyan Zhao, and Chenfei Wu. Qwen-Image-Agent: Bridging the context gap in real-world image generation, 2026. URL https://arxiv.org/abs/2606.26907.

Zeyu Zhu, Kevin Qinghong Lin, and Mike Zheng Shou. Paper2Video: Automatic video generation from scientific papers, 2025. URL https://arxiv.org/abs/2510.05096.

Zihao Zhu, Ruotong Wang, Siwei Lyu, Min Zhang, and Baoyuan Wu. BrandFusion: A multi-agent framework for seamless brand integration in text-to-video generation, 2026. URL https://arxiv.org/abs/2603.02816.