COMIC: AGENTIC SKETCH COMEDY GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a fully automated AI system that produces short comedic videos similar to sketch shows such as Saturday Night Live. Starting from character references, the system employs a population of agents loosely modeled on roles in real production studios and structured to optimize the quality and diversity of ideas and outputs through iterative competition, evaluation, and refinement. A key contribution is the introduction of LLM-based critics aligned with real viewer preferences through the analysis of a corpus of comedy videos on YouTube, enabling automatic evaluation of humor. We further propose multi-island relativistic evolution for script improvement, along with script-conditioned rendering critics that perform shot- and video-level tournaments. In both human and automated evaluations, COMIC receives higher scores than agentic and video generation baselines.

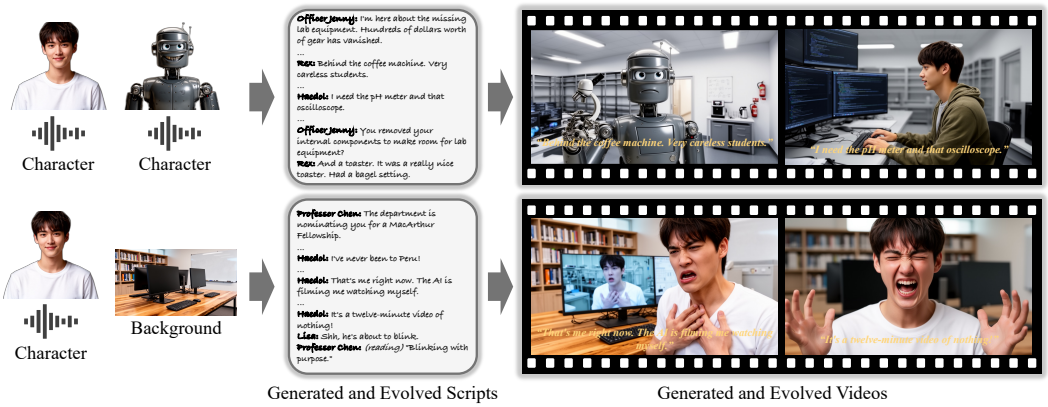


Figure 1: **COMIC** takes images, voices, and brief descriptions as input, and automatically generates entertaining comedy scripts along with corresponding video and audio. Our framework flexibly builds stories around multiple characters and custom backgrounds. Each generated comedy video is 1–2 minutes long; please watch them in the supplementary material.

1 INTRODUCTION

Can AI be funny? If you ask today's AI models to tell you a joke, you will likely receive a groan-inducing pun or a "dad joke." While generative models now excel across a wide range of writing, coding, and media generation tasks, humor remains particularly challenging. This is not to say that LLMs are incapable of humor—with the right prompts and enough iterations, one can find some gems. However, reliably producing content that can make an audience laugh is difficult.

In this paper, we propose a fully automated framework, **COMIC** (**C**ontent **O**ptimization via **M**ulti-agent **I**terative **C**ompetition), for generating short comedic videos. The input to the system is a list of character descriptions (text, image, voice) and/or background references (Figure 1). Achieving this goal requires solving three distinct tasks automatically: conceptualizing the right comedic scenario, producing a funny script, and generating a high-quality, consistent, and engaging video. Each of these tasks is challenging. Crafting genuinely funny ideas and scripts requires navigating the subjective, multidimensional space of humor, and producing long-form video remains an open problem, as state-of-the-art video models typically produce only short clips and lack strong controls for inter-clip consistency.

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.

Our approach starts from the observation that LLMs do occasionally produce humorous content when provided with the right structure. It is a bit like panning for gold—one must dig deep enough to gather sufficient material, then sift through it to find the humorous nuggets from which to build a sketch. This mirrors how real sketch comedy shows operate (Marx, 2016), with groups of writers spending many hours brainstorming and iterating before converging on a set of finalists. However, evaluating humor automatically is a big challenge. We therefore derive LLM-based *writing critics* aligned with human preferences by analyzing a corpus of YouTube comedy sketch videos and their associated viewer engagement measured by projected views.

We build a comprehensive multi-agent system loosely modeled on human production studios, using agentic versions of roles such as scriptwriters, editors, and directors, with a structure designed to encourage diversity of ideas and optimize for the emergence of humorous content. Specifically, we instantiate multiple distinct *islands* of scripts, each governed by a writing-critic committee representing a different philosophy or niche. Script populations improve on each island through round-robin tournaments in which losing scripts are refined using feedback from the winners. This topology captures the multimodal nature of humor: "good" comedy can be slapstick, dry, or surreal, and success can manifest through various approaches.

Once scripts are refined, scene director agents break each script down into distinct shots—each with its own setup, *e.g.*, characters, dialogue, expressions, and backgrounds—and render videos for each shot. Shots are produced consecutively, allowing the scene director for the current shot to reference previous shots for continuity. Each shot is evaluated by a set of script-conditioned rendering critics that embody diverse interpretations of how the narrative should be visually realized, and then refined based on critic feedback. This iterative pipeline employs depth- and breadth-wise competitions over refinement histories and scene realizations, respectively.

COMIC targets the generation of comedic videos, an open-ended creative task at the opposite extreme from mathematics or coding, which have objectively correct answers. As a systems contribution, it brings the full production workflow into one modular pipeline, from concept development and scriptwriting to storyboarding, audio-visual synthesis, continuity management, and quality control. We further demonstrate our key algorithmic contributions: generating and selecting writing critics aligned with real viewer engagement, island-based relativistic evolution for open-ended script improvement, and script-conditioned rendering critics with iterative shot- and video-level tournaments. Our experiments, across both human evaluation and independently selected benchmark video critics, show higher scores for COMIC than for the agentic and video generation baselines.

2 Background

Multi-agent evolutionary systems. Evolutionary computation has been applied to creative domains through genetic algorithms (Sims, 1991), quality-diversity methods (Mouret & Clune, 2015), and distributed island models (Tanese, 1989; Cantú-Paz et al., 1998; Whitley et al., 1999; Novikov et al., 2025). Recent LLM work applies related ideas to prompt optimization (Yang et al., 2023; Fernando et al., 2023), heuristic discovery (Liu et al., 2024), and agent improvement (Zhang et al., 2025; Yuan et al., 2025). COMIC brings this perspective to comedy, an extremely open-ended domain, through competition among multiple aligned critics.

Video generation. Text-to-video models such as Sora (OpenAI, 2025), Veo (Google DeepMind, 2025), Movie Gen (Meta, 2024), and Wan (Wan et al., 2025) generate increasingly realistic clips, while controllable systems add signals such as audio (Lin et al., 2025; Hong et al., 2025). Extensions such as StreamingT2V (Henschel et al., 2025) and FramePack (Zhang & Agrawala, 2025) increase duration, but do not address narrative coherence or comedic quality. COMIC builds a long-form production system on top of these short-form capabilities.

Agentic video production. LLM-based systems use agents for direction, planning, and self-improvement (Hong et al., 2023; Huang et al., 2023; Lin et al., 2023; Lian et al., 2023; Long et al., 2025). Vlogger (Zhuang et al., 2024) decomposes minute-level vlog production into script, actor, video, and voice stages, while Captain Cinema (Xiao et al., 2026) combines top-down keyframe planning with long-context synthesis for multi-scene movies. Storyboard-based methods (Zheng et al., 2024; Wu et al., 2025; Shi et al., 2025; Huang et al., 2025; Li et al., 2024; Dalal et al., 2025)

also target longer videos, but use shallow, single-pass workflows. COMIC instead searches the creative space through iterative competition under diverse, aligned critics.

3 Content Optimization via Multi-Agent Iterative Competition (COMIC)

3.1 Problem Statement

We address *automated sketch comedy video generation*: given character specifications $\mathcal{X} = [x_1, \ldots, x_C]$—each comprising a portrait image, voice sample, and text description—and background assets $\mathcal{B} = [b_1, \ldots, b_M]$, the system must produce a short video $\mathcal{V}^*$ that is narratively coherent, visually consistent, and funny to human viewers.

Figure 2 illustrates the two coupled subproblems. *Script generation* synthesizes $s^* \in \mathcal{S}$ through writing, comparison, and revision. *Visual realization* translates s^* into a shot sequence $\mathcal{V} = [v_1, \ldots, v_N]$ using storyboards, multimodal generation, and rendering-critic refinement.

However, a single-pass instantiation is insufficient for high-quality results; good scripts and videos are forged through multiple rounds of revision. COMIC uses human-aligned critics (Section 3.3), evolves scripts through competitive island-based search (Section 3.4), and realizes them audio-visually through iterative rendering and critic refinement (Section 3.5).

Crucially, throughout the process, we sample and leverage distinct types of critics. *Writing critics* compare and refine scripts during island evolution, while script-conditioned *rendering critics* refine and select shots and videos. *Benchmark video critics* are independently sampled and used solely for held-out automated evaluation (Section 4.3).

3.2 Why Fixed Objectives Fall Short

Traditional goal-based optimization presupposes a stationary reward function $R\colon \mathcal{S} \to \mathbb{R}$. However, humor is inherently context-dependent and subjective. A fixed scalar objective invites Goodhart's Law (Goodhart, 1984), rewarding a proxy metric rather than genuine creative quality. Furthermore, great humor takes many distinct forms; slapstick and dry wit share no common measuring stick, and different people prefer different styles. It is impossible to aggregate preference profiles into a single, consistent ranking without sacrificing desirable properties (Arrow, 1950). Meanwhile, existing agentic video systems (Lin et al., 2023; Zheng et al., 2024; Wu et al., 2025) rely on fixed roles in a shallow pipeline, limiting feedback and alternative perspectives. In contrast, COMIC introduces multiple relativistic islands, each with its own comparative objective, perspective, and niche. Within the framework, a script's fitness is evaluated relative to current competitors on the same island, and losing scripts are constantly refined using critic feedback. Consequently, the multidimensional competitive baseline can continuously rise without requiring a fixed destination.

3.3 Alignment of Writing Critics to Real Viewers

Writing quality depends critically on writing-critic selection. Rather than hand-crafting prompts or fine-tuning a dedicated evaluator, we use a *generate-and-select* strategy based on a large corpus of videos. We synthesize a diverse pool of candidate writing critics, each defined by a system prompt specifying its persona, and retain those whose preferences align with empirical audience engagement signals. This provides diverse evaluative perspectives while enabling aggressive pruning to retain only informative critics.

Engagement scoring. We collect 4,940 data points from five YouTube sketch comedy channels: Foil Arms & Hog, Key & Peele, SNL, Studio C, and Viva La Dirt League. We use view counts as a proxy for popularity. Since view-count trajectories empirically follow an S-curve, we fit a per-channel logistic growth model, $V(t) = L/(1 + \exp(-r(t - t_0)))$, by nonlinear least squares. We define each video's *engagement score* as its projected infinite-horizon view count, $L_{\text{proj}} = V(t) \cdot (1 + \exp(-r(t - t_0)))$. Scripts are then selected for top, middle, and bottom engagement tiers and partitioned into $\mathcal{S}_{\text{in-context}}$, $\mathcal{S}_{\text{val}}$, and held-out $\mathcal{S}_{\text{test}}$. The per-channel fitted trajectories are shown in Appendix B, and additional processing details are provided there.

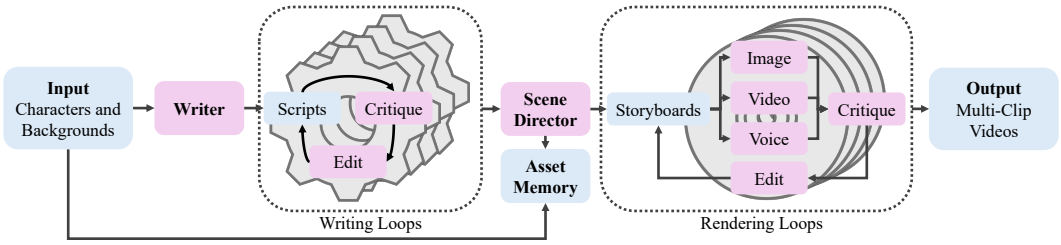


Figure 2: Overall agentic flow. COMIC is loosely modeled on a production studio, with agentic counterparts for a writer, critic, editor, and scene director. The writing and rendering loops generate scripts and videos through island-based competition and iteration.

Writing-critic pool generation. We construct a diverse pool $\mathcal{C}_{\text{pool}}$ by prompting a meta-critic p_{script} with stratified, tier-labeled examples:

$$\mathcal{C}_{\text{pool}} \sim p_{\text{script}}(\mathcal{C} \mid \mathcal{S}_{\text{in-context}}), \tag{1}$$

This calibrates each writing critic's aesthetic preferences to channel-specific engagement patterns. Rather than producing a single critic, we sample a set of critics (10 in practice) and select the subset that best discriminates real viewer engagement scores.

Task-specific selection. To capture both coarse and fine quality distinctions, we define two pairwise tasks. *Top vs. Bottom* targets critics sensitive to large quality gaps, while *Top vs. Middle* targets critics sensitive to subtle distinctions. For each channel χ and sensitivity level τ, we select the highest-accuracy critic on validation task $\mathcal{T}^{\text{val}}_{\chi,\tau}$:

$$c^*_{\chi,\tau} = \underset{c \in \mathcal{C}_{\text{pool}}}{\operatorname{argmax}} \operatorname{Acc}(c \mid \mathcal{T}^{\text{val}}_{\chi,\tau}), \tag{2}$$

This yields the specialized pool $\mathcal{C}_{\text{task}} = \bigcup_{\chi,\tau} \{c^*_{\chi,\tau}\}$. To evaluate whether this critic pool is well-aligned and diverse to represent a broader pool, we compare it against both the ensemble mean of all critics and the top-performing single-critic pool.

Notably, task-specific selection improves held-out accuracy over the mean and single-best baselines (Table 1). We provide channel-specific validation results in Table 6 of Appendix B, which also follow a similar trend.

Table 1: Held-out test accuracy for writing critics.

Comparison	Mean Critic	Single Best	Task-Wise Best
Top vs. Middle	0.557	0.554	**0.578**
Top vs. Bottom	0.654	0.670	**0.716**

In Table 2, we demonstrate the impact of in-context sample size on the performance of selected writing critics. Specifically, we analyze how the number of tier-labeled samples provided to p_{script} affects its ability to yield high-performing critics by evaluating settings with 0 (zero-shot), 15, and 45 samples across the selection strategies. Although even the zero-shot setting achieves favorable accuracy, providing 45 samples yields the best overall performance. Note that across all sample sizes, *Task-Wise Best* remains the strongest approach.

Table 2: Writing-critic validation accuracy as a function of in-context sample size. Task-wise selection has the highest reported accuracy for each sample size.

# Samples	Comparison	Mean Critic	Single Best	Task-Wise Best
0	Top vs. Middle	0.542	0.572	0.642
	Top vs. Bottom	0.700	0.728	0.802
15	Top vs. Middle	0.547	0.590	0.644
	Top vs. Bottom	0.697	0.702	0.808
45	Top vs. Middle	0.542	0.572	0.644
	Top vs. Bottom	0.708	0.740	0.830

3.4 Script Writing Loop

Islands and evolving fitness landscapes. We introduce an approach to iteratively evolve a population of scripts. Pairs of scripts are compared by a writing critic, which provides feedback to revise

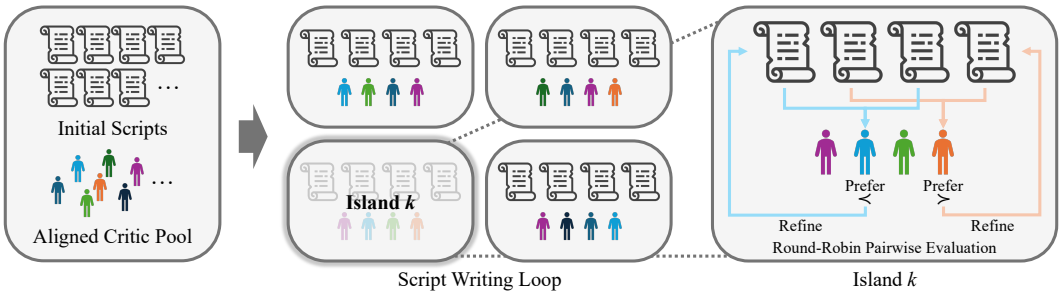


Figure 3: Script writing loop. Isolated script populations evolve on separate islands under distinct writing-critic committees sampled from the aligned pool. Losing scripts are refined through round-robin pairwise tournaments.

the weaker script. As the population evolves, weaker scripts are iteratively refined using the feedback, continuously raising the competitive baseline. A script that wins at generation g may lose at $g+1$ not because it degraded, but because competitors improved. Defining fitness $f^{(g)}(s)$ as the expected win rate of script s against the current population given a writing-critic committee $\mathcal{C}$ and script population at the current generation $\mathcal{S}^{(g)}$,

$$f^{(g)}(s) = \mathbb{E}_{s' \sim \mathcal{S}^{(g)}, c \sim \mathcal{C}} \big[\mathbb{I}\big[c(s, s') \mapsto (s, \cdot) \big] \big], \tag{3}$$

this formulation ideally implements a competitive environment that grows more demanding, *i.e.*, $\mathbb{E}[f^{(g)}(s)] \geq \mathbb{E}[f^{(g+1)}(s)]$, forcing continuous adaptation.

To encourage diversity of solutions, we partition the global script population into K isolated islands $\{I_1, \ldots, I_K\}$, each governed by a specialized writing-critic committee $\mathcal{C}_k$ drawn from $\mathcal{C}_{\text{task}}$ in Section 3.3. These separate committees embody distinct comedic preferences, while being aligned with engagement patterns. The fitness landscape on island k is shaped by two coupled elements: (1) the island-specific writing-critic committee $\mathcal{C}_k$, which defines evaluative standards, and (2) the evolving script population $\mathcal{S}_k$, which determines the comparative baseline. Because both critics and populations differ across islands, the fitness landscapes tend to diverge, yielding a Pareto frontier of diverse comedic styles. Figure 3 illustrates this island structure and the critic-guided competition.

Pairwise evaluation with critic-guided update. Within each island, evolution proceeds through round-robin pairwise evaluation. At each iteration, two scripts $s_i, s_j \in \mathcal{S}_k$ are compared by every critic in the local committee. Each critic $c_e \in \mathcal{C}_k$ performs an independent evaluation:

$$c_e(s_i, s_j) \;\mapsto\; (w_{c_e},\, \phi_{c_e}). \tag{4}$$

Let $s_\ell \in \{s_i, s_j\} \setminus \{w_{c_e}\}$ denote the losing script. It undergoes a *critic-guided update* driven by the feedback:

$$s_\ell \;\leftarrow\; U(s_\ell,\, \phi_{c_e}), \tag{5}$$

where $U : \mathcal{S} \times \Phi \to \mathcal{S}$ is the update operator that rewrites the script according to natural-language feedback ϕ_{c_e}, $\mathcal{S}$ denotes the space of scripts, and Φ denotes the space of natural-language feedback. Note that we only update the losing script. This compact operator integrates two classical evolutionary mechanisms in a single call: the comparative feedback ϕ_{c_e} encourages the loser to incorporate the winner's strengths, resulting in semantic *crossover* that transfers beneficial features from the superior script, while U simultaneously introduces semantic *mutation* by rewriting the script under critic guidance, exploring variations that may uncover novel comedic approaches.

3.5 Video Rendering Loop

The rendering stage translates selected scripts into videos using script-conditioned critics and competition-based refinement. Figure 4 shows a rough overview of this process.

Script-conditioned rendering-critic generation. Given a refined script s, a video meta-critic p_{render} generates a script-conditioned set of rendering critics:

$$\mathcal{C}_{\text{render}} \sim p_{\text{render}}(\mathcal{C} \mid s) \tag{6}$$

Each critic provides a different lens for visually realizing the narrative.

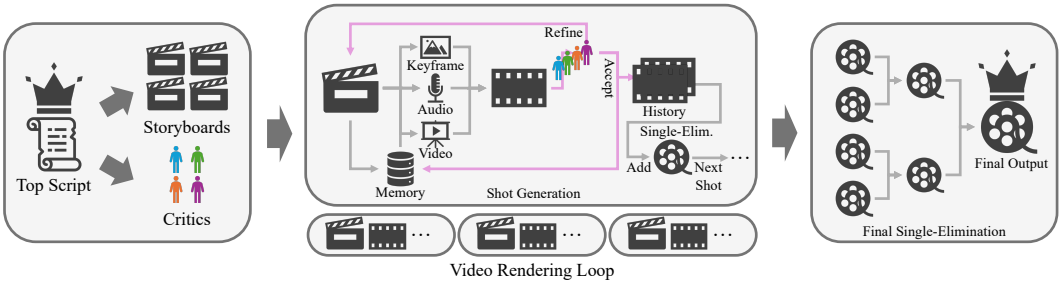


Figure 4: Video rendering loop. Script-conditioned rendering critics refine scene directions, while shot- and video-level tournaments select from refinement histories and scene realizations, *i.e.*, depth and breadth, respectively.

Storyboarding. Because video generation is computationally expensive, we first storyboard the visual content conditioned on the script. For each script s, a scene director generates D scene directions $\{d^{(1)}, \ldots, d^{(D)}\}$, where each $d^{(j)} = \{d_1^{(j)}, \ldots, d_{N_j}^{(j)}\}$ specifies N_j shots in text form. Each shot specification defines reference characters, backgrounds, and previous shots, along with poses, expressions, camera framing, or actions. A structured memory bank $\mathcal{M}$ stores character assets, backgrounds, and the last frame of each finalized shot so subsequent specifications can preserve continuity. In addition, we use chain-of-thought reasoning by introducing setup notes to help the director plan complex visual arrangements before producing shot specifications.

Iterative shot refinement with history tournament. For direction $d^{(j)}$ and shot i, we render an initial shot and iteratively revise its specification:

$$v_i^{(0)} \leftarrow \text{Render}\big(d_i^{(j,0)}, V_{<i}, \mathcal{M}\big), \tag{7}$$

where Render includes image, voice, and video generation based on the direction. For $m = 0, \ldots, |\mathcal{C}_{\text{render}}| - 1$, a rendering critic evaluates the current shot and proposes a refined specification:

$$c\big(d_i^{(j,m)}, V_{<i}, v_i^{(m)} \mid s\big) \mapsto \big(d_i^{(j,m+1)}, \phi_{\text{refine}}\big), \tag{8}$$

where $V_{<i} = [v_1, \ldots, v_{i-1}]$ are previously finalized shots. The updated specification sequentially guides the next render.

Refinement accumulates a history $\mathcal{H}_i^{(j)} = \{v_i^{(0)}, \ldots, v_i^{(|\mathcal{C}_{\text{render}}|)}\}$. Rather than accepting the final iteration, we tournament the refinement history $\mathcal{H}_i^{(j)} = \{v_i^{(0)}, \ldots, v_i^{(|\mathcal{C}_{\text{render}}|)}\}$ and select the winner:

$$v_i^* = \text{SingleElimination}\big(\mathcal{H}_i^{(j)}, \mathcal{C}_{\text{render}}, s, V_{<i}\big). \tag{9}$$

This guards quality against degradation after refinement. The history tournament selects with respect to *depth*, encouraging exploitation of the given scene direction.

Test-time scaling via scene-level tournament. After rendering the D scene directions into complete videos $\{\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(D)}\}$, where each $\mathcal{V}^{(j)} = [v_1^*, \ldots, v_{N_j}^*]$ consists of selected shots, we perform a final tournament:

$$\mathcal{V}^* = \text{SingleElimination}\big(\{\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(D)}\}, \mathcal{C}_{\text{render}}, s\big). \tag{10}$$

Scaling D at inference time directs additional compute toward rendering quality without retraining, providing a parallelizable test-time scaling axis. The scene-level tournament selects across *breadth*, encouraging exploration of diverse scene realizations.

4 Experiments

4.1 Implementation Details

COMIC exposes several scaling dimensions: the number of islands K, scripts per island $|\mathcal{S}_k|$, writing critics per island $|\mathcal{C}_k|$, scene directions D, and rendering critics $|\mathcal{C}_{\text{render}}|$. We define small, base,

Table 3: Human evaluation of generated videos. Scores are means on 7-point scales. *vs. Human* compares each average, non-curated output with professional, human-made sketch comedy.

Method	Funniness↑	Watch More↑	vs. Human↑	Script↑	Narrative↑	Realism↑	Consistency↑
Veo 3.1 (Google DeepMind, 2025)	2.32	2.36	2.27	2.18	3.32	4.91	5.05
Sora 2 (OpenAI, 2025)	2.73	2.73	2.32	2.45	3.36	5.73	5.50
VGoT (Zheng et al., 2024)	1.18	1.27	1.14	1.00	1.23	2.00	2.32
MovieAgent (Wu et al., 2025)	1.27	1.09	1.18	1.09	1.09	1.27	1.14
COMIC (Ours)	**3.45**	**3.09**	**3.05**	**3.32**	**4.50**	**4.27**	**4.50**

Table 4: Automated evaluation by benchmark video critics independently sampled and separately selected from COMIC's optimization critics, averaged across channels. *Single Best* uses one top benchmark critic; *Channel-Wise Best* aggregates per-channel benchmark critics.

Method	Single Best			Channel-Wise Best		
	Win Rate	Inter-Diversity	Intra-Diversity	Win Rate	Inter-Diversity	Intra-Diversity
Veo 3.1 (Google DeepMind, 2025)	0.010	0.308	0.369	0.105	0.263	0.360
Sora 2 (OpenAI, 2025)	<u>0.075</u>	<u>0.531</u>	**0.722**	<u>0.175</u>	<u>0.310</u>	<u>0.563</u>
VGoT (Zheng et al., 2024)	0.000	0.000	0.000	0.010	0.105	0.189
MovieAgent (Wu et al., 2025)	0.000	0.000	0.000	0.130	0.088	0.180
COMIC (Ours)	**0.440**	**0.780**	<u>0.682</u>	**0.390**	**0.519**	**0.693**

and large configurations across these dimensions. Unless otherwise noted, we report results from the 4th generation and use the base configuration. The base configuration runs in approximately one day on a single GPU with an API budget of around $5. Model details and full configurations are in Appendix C (Tables 7 and 8).

We evaluate scripts and videos by pairwise comparisons. Let A, B, and E denote reference samples, generated samples, and evaluators, and let $P_{e,b,a}$ be the probability that b beats a under e. *Win Rate* measures average quality, $Q_{\text{avg}} = \mathbb{E}_{e,b,a}[P_{e,b,a}]$, where values above 0.5 favor generated samples. *Inter-Diversity* measures variation across generated samples, $D_{\text{inter}} = \mathbb{E}_{e,a}[\text{Var}_b(P_{e,b,a})]/(Q_{\text{avg}}(1-Q_{\text{avg}}))$. *Intra-Diversity* measures evaluator/reference variation for each generated sample, $D_{\text{intra}} = \mathbb{E}_b[\text{Var}_{e,a}(P_{e,b,a})]/(Q_{\text{avg}}(1-Q_{\text{avg}}))$. The normalization makes both diversity metrics scale-invariant.

4.2 Video Results

The Figure 1 teaser illustrates sketch comedy videos generated by COMIC from minimal specifications (*e.g.*, a portrait, voice sample, and brief description). Starting from these inputs, the system autonomously develops comedic arcs with setups, escalating tension, and payoffs. The sketches span dry, deadpan, and surreal styles while maintaining character, background, and shot continuity; additional videos are provided in Figure 10 in Appendix A, and the supplementary material.

4.3 Baseline Comparison

We compare COMIC against VideoGen-of-Thought (Zheng et al., 2024) and MovieAgent (Wu et al., 2025) as agentic video production baselines. These represent storyboard-driven long-form agentic video generation but lack iterative refinement or competitive selection. We further compare against black-box frontier models Sora 2 (OpenAI, 2025) and Veo 3.1 (Google DeepMind, 2025), to assess the contribution of our agentic pipeline over raw generative capability.

Automated evaluation. For benchmarking, we prompt a video meta-critic agent to synthesize a pool of candidate benchmark critics with diverse personas. Crucially, this pool is independently sampled and its critics are separately selected from the writing and script-conditioned rendering critics. The selected benchmark critics compare generated videos with held-out middle-tier test videos representing "median" sketch comedies.

We consider two aggregation approaches: *Single Best*, which selects the highest-accuracy critic on validation, and *Channel-Wise Best*, which selects critics independently per channel to capture diverse comedic traditions. Table 4 reports win rate, inter-diversity, and intra-diversity averaged across channels. COMIC has the highest win rate under both aggregations, and also has the highest inter- and intra-diversity; the automated ordering (COMIC $>$ Sora $>$ Veo $>$ MA $\approx$ VGoT) agrees with the human evaluation results presented in the following paragraph.

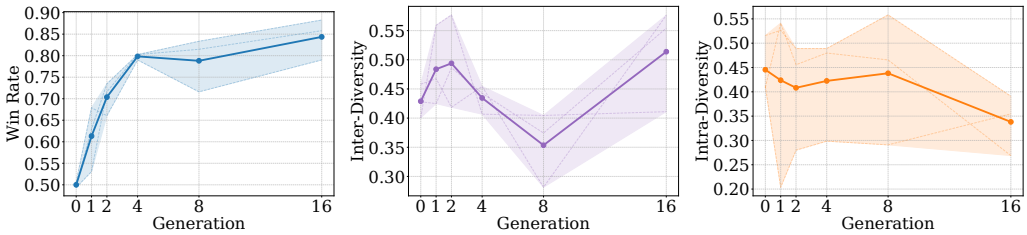
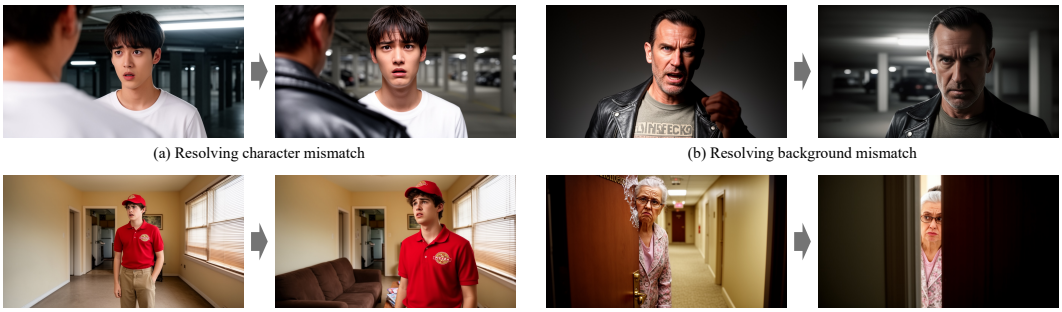


Figure 5: Effect of generation, measured by *Win Rate*, *Inter-Diversity*, and *Intra-Diversity* across iterations relative to the 0th generation. Average values are presented as solid lines, and the range is depicted as shading. Through the 4th generation, the win rate increases substantially. *Inter-Diversity* (diversity across scripts) initially drops due to the emergence of coherent, favorable responses but increases as generations progress, driven by our divergent mechanism.



(a) Resolving character mismatch

(b) Resolving background mismatch

(c) Adjusting framing and props

(d) Fixing visual artifacts

Figure 6: Qualitative examples of corrections made by script-conditioned rendering critics, including character, background, framing, and visual-artifact fixes.

Human evaluation. We conducted a blind, randomized human evaluation of five methods using automatically selected, non-curated COMIC outputs. Table 3 reports mean scores on 7-point scales. COMIC receives the highest mean scores among the agentic methods across *Funniness*, *Watch More*, *Script*, *Narrative*, *Realism*, and *Consistency*. Although Sora 2 and Veo 3.1 score higher on *Realism* and *Consistency*, COMIC scores higher on *Funniness*, *Watch More*, *vs. Human*, *Script*, and *Narrative* while producing longer videos. For further details and the evaluation protocol, see Appendix H.

Comparison against human-produced content. On the *vs. Human* dimension, COMIC's average non-curated outputs score 3.05, while the best agentic baseline achieves only 1.14. The reference point of 4 is intentionally demanding, representing the level of typical professional comedy shows that audiences consume and serving as our North Star rather than a mediocre AI generation baseline. Showing meaningful progress, COMIC comes closer to this point than the frontier video models and agentic baselines.

Controlled backbones and refinement. We rerun VGoT and MovieAgent with the language, image, and video backbones used by COMIC. As shown in Table 5, our full pipeline consistently outperforms the controlled baselines in win rate and inter-diversity across both sets of benchmark critics.

Table 5: Controlled-backbone automated evaluation. Starred rows indicate COMIC backbones. *No-Refine* is an ablation of our framework that removes multi-island iterative refinement.

Benchmark critic	Metric	VGoT*	MovieAgent*	No-Refine*	**COMIC (Full)***
Single Best	Win Rate↑	0.03	0.02	0.30	**0.44**
	Inter-Div.↑	0.62	0.46	0.70	**0.78**
Channel-Wise Best	Win Rate↑	0.17	0.13	0.31	**0.39**
	Inter-Div.↑	0.37	0.33	0.50	**0.52**

4.4 Ablation Study

Island-based evolution and iterative refinement. Figure 5 tracks win rates and diversity across successive generations. The win rate rises sharply through generation 4, confirming that pairwise tournaments drive rapid improvement. Inter-population diversity initially drops as populations converge toward generally effective strategies, then recovers as distinct critic committees push them to-

ward unique niches. Examples of script refinement are shown in Figure 12 in Appendix E. Figure 6 illustrates how rendering critics correct issues such as character mismatches and framing errors.

Single pool vs. multi-island. We compare a single unified script pool with the multi-island topology while keeping the number of refinement iterations per script fixed. In the single-pool setting, every script is optimized under one shared population and selection pressure. In contrast, islands maintain separate populations and distinct writing-critic committees during local tournaments, allowing different comedic directions to develop before the final selection. As shown in Figure 7, the multi-island configuration improves both win rate and intra-diversity. Thus, preserving multiple local selection landscapes retains meaningfully distinct script trajectories without compromising average quality.

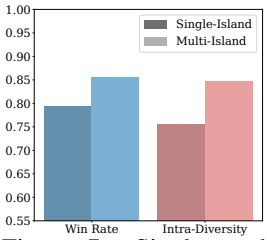


Figure 7: Single- and multi-island settings.

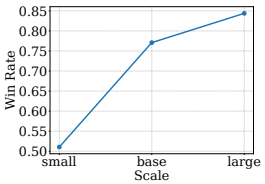


Figure 8: Scale.

Scale. We compare small, base, and large configurations by jointly scaling the number of islands, scripts, and critics. Fig. 8 shows the win rate of each relative to the small scale. Performance improves consistently as the system is scaled. Top scripts from the large configuration achieve a higher win rate compared to the small and base baselines, showing that COMIC scales by trading test-time compute for enhanced performance. The full configurations are listed in Appendix D, Table 8.

No-critics baseline. Fig. 9 reports an A/B preference study comparing the full COMIC pipeline with a critic-free ablation. Raters preferred the full system on *Script*, *Narrative*, *Realism*, *Consistency*, and *Funniness*. The largest separations occur on visual consistency and realism, which are directly targeted by the script-conditioned rendering-critic loop, and the writing-oriented metrics also favor the full system. Additionally, automated results in Table 5 show that COMIC also outperforms *No-Refine* in both win rate and inter-diversity.

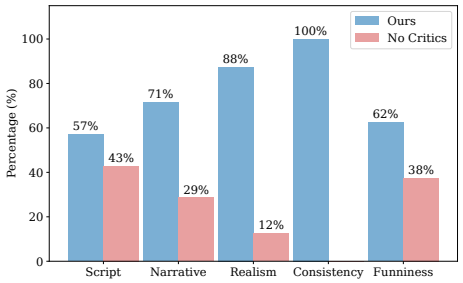


Figure 9: Human preference study against the critic-free ablation.

5 Conclusion

In this paper, we introduce COMIC, a fully automated multi-agent framework designed to tackle the extremely open-ended challenge of sketch comedy video generation. By shifting away from single-pass, fixed-objective pipelines, COMIC leverages a multi-island topology where diverse, human-aligned critic committees drive iterative refinement. The competitive pressure operates across both the narrative scriptwriting and visual rendering stages, allowing the system to explore a vast creative space while maintaining coherence. Our experiments demonstrate that COMIC outperforms existing agentic video baselines while offering a dual mechanism for test-time scaling. Ultimately, this work establishes a new state of the art for automated, engaging, long-form video production.

The iterative refinement process incurs computational costs, and normalized YouTube view counts can be noisy proxies for humor quality. Future work could add sound effects and quantify originality and attribution when building on large internet corpora.

COMIC's improvements emerge without parameter updates, gradient-based optimization, or a fixed reward signal, connecting to the *Red Queen* hypothesis (Van Valen, 1973) in evolutionary biology, wherein species must continuously evolve to maintain their fitness against co-evolving competitors. Unlike structured domains such as mathematics (Woodruff et al., 2026) or board games (Silver et al., 2016), comedy's shifting, context-dependent criteria make it a compelling proxy for open-ended, real-world problems. We believe that this work opens several directions for future research into other creative domains.

REFERENCES

Kenneth J Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4): 328–346, 1950.

Erick Cantú-Paz et al. A survey of parallel genetic algorithms. *Calculateurs paralleles, reseaux et systems repartis*, 10(2):141–171, 1998.

Hyunmin Cho, Donghoon Ahn, Susung Hong, Jee Eun Kim, Seungryong Kim, and Kyong Hwan Jin. Tag: Tangential amplifying guidance for hallucination-resistant diffusion sampling. *arXiv preprint arXiv:2510.04533*, 2025.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17702–17711, 2025.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

Charles AE Goodhart. Problems of monetary management: the uk experience. In *Monetary theory and practice: The UK experience*, pp. 91–121. Springer, 1984.

Google DeepMind. Veo 3.1. https://deepmind.google/models/veo/, 2025.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2568–2577, 2025.

Susung Hong, Junyoung Seo, Heeseong Shin, Sunghwan Hong, and Seungryong Kim. Direct2v: Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint arXiv:2305.14330*, 2023.

Susung Hong, Ira Kemelmacher-Shlizerman, Brian Curless, and Steven M Seitz. Musicinfuser: Making video diffusion listen and dance. *arXiv preprint arXiv:2503.14505*, 2025.

Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *Advances in Neural Information Processing Systems*, 36:26135–26158, 2023.

Kaiyi Huang, Yukun Huang, Xintao Wang, Zinan Lin, Xuefei Ning, Pengfei Wan, Di Zhang, Yu Wang, and Xihui Liu. Filmaster: Bridging cinematic principles and generative ai for automated film generation. *arXiv preprint arXiv:2506.18899*, 2025.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Yunxin Li, Haoyuan Shi, Baotian Hu, Longyue Wang, Jiashun Zhu, Jinyi Xu, Zhen Zhao, and Min Zhang. Anim-director: A large multimodal model powered agent for controllable animation video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13847–13858, 2025.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023.

Fei Liu, Xialiang Tong, Mingxuan Yuan, Xi Lin, Fu Luo, Zhenkun Wang, Zhichao Lu, and Qingfu Zhang. Evolution of heuristics: Towards efficient automatic algorithm design using large language model. *arXiv preprint arXiv:2401.02051*, 2024.

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.

Nick Marx. Live from new york: The complete uncensored history of "saturday night live" as told by its stars, writers, and guests, 2016.

Meta. Meta movie gen. https://ai.meta.com/research/movie-gen/, 2024.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

OpenAI. Sora 2. https://openai.com/index/sora-2/, 2025.

Resemble AI. Chatterbox-TTS. `https://github.com/resemble-ai/chatterbox`, 2025. GitHub repository.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. Animaker: Automated multi-agent animated storytelling with mcts-driven clip generation. *arXiv preprint arXiv:2506.10540*, 2025.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Karl Sims. Artificial evolution for computer graphics. In *Proceedings of the 18th annual conference on Computer graphics and interactive techniques*, pp. 319–328, 1991.

Reiko Tanese. *Distributed genetic algorithms for function optimization*. University of Michigan, 1989.

Leigh Van Valen. A new evolutionary law. *Evolutionary theory*, 1:1–30, 1973.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Darrell Whitley, Soraya Rana, and Robert B Heckendorn. The island model genetic algorithm: On separability, population size and convergence. *Journal of computing and information technology*, 7(1):33–47, 1999.

David P Woodruff, Vincent Cohen-Addad, Lalit Jain, Jieming Mao, Song Zuo, MohammadHossein Bateni, Simina Branzei, Michael P Brenner, Lin Chen, Ying Feng, et al. Accelerating scientific research with gemini: Case studies and common techniques. *arXiv preprint arXiv:2602.03837*, 2026.

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

Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan Yuille, and Lu Jiang. Captain cinema: Towards short movie generation. In *International Conference on Learning Representations*, 2026.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2023.

Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Dongsheng Li, and Deqing Yang. Evoagent: Towards automatic multi-agent generation via evolutionary algorithms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6192–6217, 2025.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin godel machine: Open-ended evolution of self-improving agents. *arXiv preprint arXiv:2505.22954*, 2025.

Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, et al. Videogen-of-thought: Step-by-step generating multi-shot video with minimal manual intervention. *arXiv preprint arXiv:2412.02259*, 2024.

Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Appendix

A Video Results

We include the MP4 video files as accompanying files and encourage readers to watch them.



Figure 10: Additional sketch comedy videos featuring varied generated situations.

B Writing-Critic Selection

Channel-specific validation. Table 6 reports writing-critic validation accuracy by channel and engagement tier, complementing the in-context sample-size comparison in the main paper (Table 2). Task-wise selection achieves the highest average accuracy for both comparisons.

Table 6: Channel-specific writing-critic validation accuracy. Task-wise selection has the highest average accuracy across the reported channels and engagement tiers. FAH = Foil Arms & Hog; VLDL = Viva La Dirt League; K&P = Key & Peele.

	Comparison	Studio C	FAH	VLDL	K&P	SNL	**Average**
Mean Critic	Top vs. Middle	0.63	0.40	0.58	0.53	0.59	0.55
	Top vs. Bottom	0.72	0.62	0.79	0.60	0.81	0.71
Single Best	Top vs. Middle	0.62	**0.52**	0.56	0.57	0.59	0.57
	Top vs. Bottom	0.67	0.70	0.84	0.69	0.80	0.74
Task-Wise Best	Top vs. Middle	**0.70**	**0.52**	**0.70**	**0.65**	**0.65**	**0.64**
	Top vs. Bottom	**0.85**	**0.78**	**0.85**	**0.72**	**0.95**	**0.83**

Data processing details. We collect 4,940 data points from five YouTube sketch comedy channels, excluding videos that do not meet our criteria for length and format. To construct channel-specific engagement scores, we model the cumulative view trajectory of each sketch using a logistic growth function. This reflects the S-shaped trajectories observed in online video view counts. Figure 11

presents the fits of the logistic growth model for all five channels. Based on the projected infinite-horizon view count L_{proj}, we extract 30 data points from each of the five channels and each of the three tiers, resulting in 450 total data points. We then split them into $\mathcal{S}_{\text{in-context}}$, $\mathcal{S}_{\text{val}}$, and $\mathcal{S}_{\text{test}}$ as mentioned in the main paper.

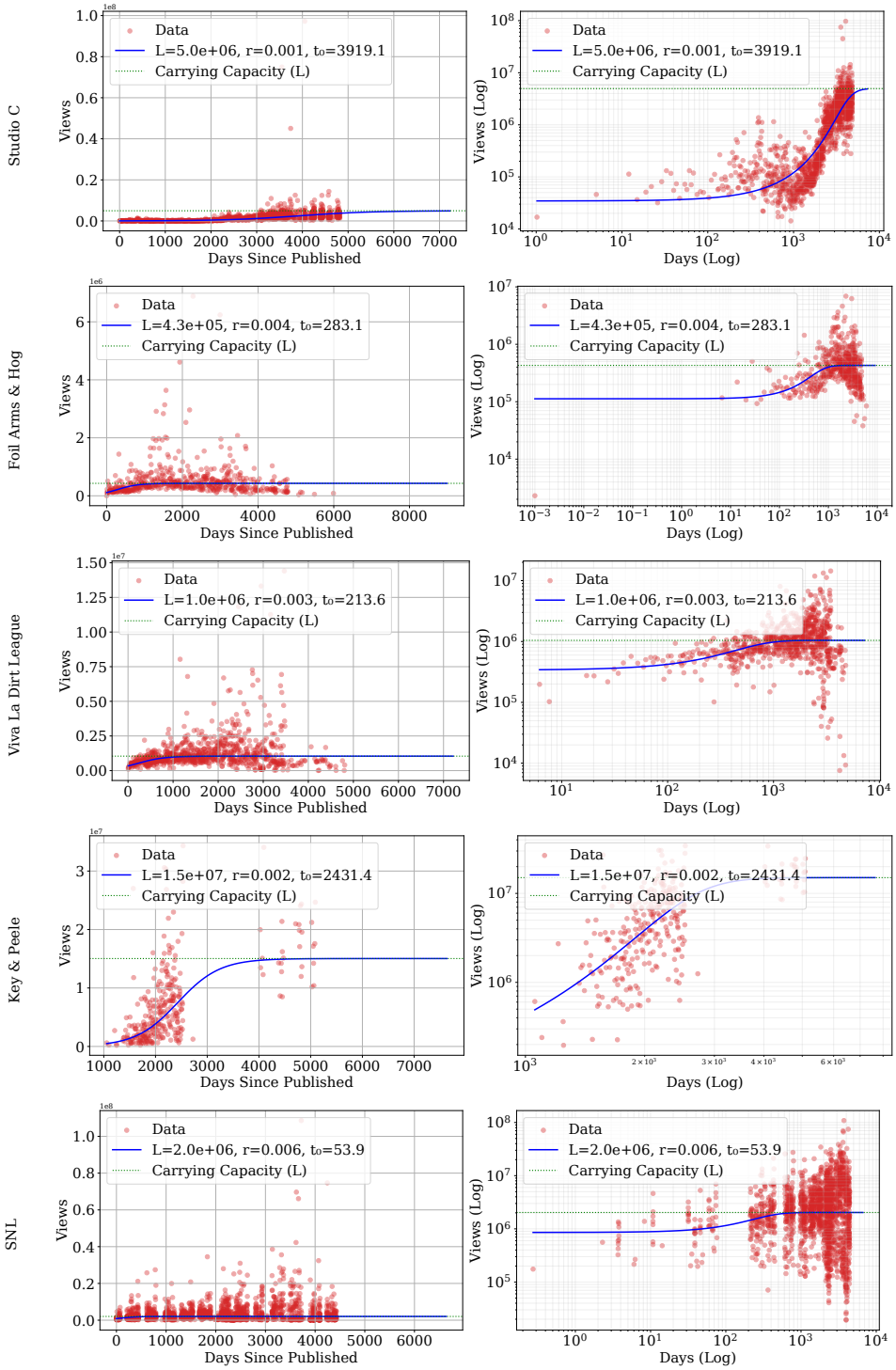


Figure 11: Logistic growth model fits for cumulative view counts across different comedy channels.

C MODELS

A key feature of our framework is the ability to easily integrate modular foundation models at different production stages. Table 7 summarizes the models used throughout the COMIC pipeline. We use Claude 4 Sonnet for concept and script generation. For tasks requiring efficiency during iterative refinement, we use Claude 4.5 Haiku for language-only writing critics and Gemini 3 Flash Preview for multimodal rendering and benchmark video critics. For tasks requiring more robust reasoning—such as script inspection, meta-critic instruction generation, and final scene direction—we employ Claude 4.5 Opus. We also use this model for the controlled backbone ablation.

For visual synthesis, we use FLUX.2 [dev] (Labs, 2025) to generate canonical character appearances and enhance visual clarity. For voice consistency, we employ ElevenLabs and Chatterbox-TTS (Resemble AI, 2025) to generate stable voice prints for each character. These assets are stored in the visual memory bank $\mathcal{M}_{\text{visual}}$ and retrieved during shot generation. Video rendering leverages Wan 2.2 (Wan et al., 2025). Since our image and video generation requires taking various modalities of input, *e.g.*, image, text, and audio, we add a manifold-tangential guidance term to the predicted denoising score (Cho et al., 2025), which is agnostic to the input condition yet improves visual clarity and reduces hallucination. Additionally, we use an image critic to perform best-of-batch selection to further enhance quality and consistency. All image, video, and voice generation is performed on a single H200.

Table 7: Models used in the pipeline.

Component	**Model**
Concept and Script Writing	Claude 4 Sonnet
Language-Only Writing Critics	Claude 4.5 Haiku
Rendering / Benchmark Video Critics	Gemini 3 Flash Preview
Script Inspection	Claude 4.5 Opus
Scene Direction Generation	Claude 4.5 Opus
Meta-Critic	Claude 4.5 Opus
Character Image Synthesis	FLUX.2 [dev] + TAG
Voice Synthesis	ElevenLabs
Voice Cloning	Chatterbox-TTS
Video Rendering	Wan 2.2 + TAG

D SCALE CONFIGURATIONS

COMIC exposes several useful scaling dimensions. Round-robin pairwise evaluation within islands requires $\mathcal{O}(K|\mathcal{S}_k|^2|\mathcal{C}_k|)$ writing-critic calls per generation, which is substantially lower than a global tournament over all scripts that would require $\mathcal{O}(K^2|\mathcal{S}_k|^2|\mathcal{C}_k|)$ calls. Island-local evaluation is, therefore, not merely a diversity mechanism but also a computational necessity that makes iterative refinement at scale tractable. Table 8 summarizes the three scale configurations of COMIC.

Table 8: Scale configurations for the COMIC framework.

Scale	K	$\lvert S_k\rvert$	$\lvert\mathcal{C}_k\rvert$	D	$\lvert\mathcal{C}_{\text{render}}\rvert$
Small	2	2	2	1	0
Base	3	3	3	2	1
Large	4	4	4	4	2

Computational complexity. We analyze the computational cost of COMIC across its key scaling dimensions: K islands, $|S_k|$ scripts per island, $|\mathcal{C}_k|$ writing critics per island, D scene directions, $|\mathcal{C}_{\text{render}}|$ rendering critics, maximum $N := \max_j N_j$ shots per video, and G generations.

Within each island, a single generation requires a round-robin pairwise tournament over $|S_k|$ scripts, each evaluated by all $|\mathcal{C}_k|$ writing critics, incurring $\mathcal{O}(|S_k|^2|\mathcal{C}_k|)$ evaluations per island. Across K

Script 1 snippet

Melissa Swift: Darling, this collection is revolutionary! The raw authenticity!
Haedol Park: Thanks, I guess? I just grab whatever's free at conferences.
Marcus Webb: Are we seriously doing a fashion shoot with a database shirt?
~~Melissa Swift: Not databases, Marcus. Data liberation! Now drape it like liquid mercury.~~
Melissa Swift: Not databases, Marcus. Data liberation! This one's our cover piece.
Haedol Park: It's got a stain on it.
Melissa Swift: Vintage patina! We'll charge extra for authenticity!
Marcus Webb: That's marinara sauce.
Melissa Swift: Even better. Organic, handcrafted details. Now drape it like liquid mercury.

Script 2 snippet

Gloria: Your file is not red. No one has a red file.
~~Eric Cross: One blink. Confirmed red. I'm going to need a new identity and a vehicle registered under an alias.~~
Eric Cross: One blink. Confirmed red. I need to see the DMV director.
~~Gloria: We can't do that.~~
Gloria: There is no director.
~~Eric Cross: Of course you can't. You're not authorized. Who is? Who's your supervisor?~~
Eric Cross: Exactly. That's what the director would say. What's behind that door?

Script 3 snippet

~~Campus Security: They're free. I get them at every event. I have eighty in my office. I use them for car wash rags.~~
Campus Security: They're free. I get them at every event. I have eighty in my office.
The Shirt Boss: You attend conferences?
Campus Security: No, I work the registration tables. They give them to me in bulk. I use them for car wash rags.
The Shirt Boss: [staring] Car wash rags?
~~Campus Security: The 2019 DeepMind one's particularly absorbent. You want some? [holds out a pile] I've got a whole filing cabinet.~~
Campus Security: The 2019 DeepMind one's particularly absorbent. You want some? [holds out a pile] I've got a whole filing cabinet. There's also a stack in the lost and found from 2017. Nobody claimed them.
~~[The Shirt Boss and Nervous Ned exit silently]~~
[The Shirt Boss and Nervous Ned exit silently, defeated]
Haedol Park: What about the DeepMind?
~~Campus Security: That one? Yeah, take it. Already used it twice.~~
Campus Security: That one? Yeah, take it. Already used it twice. [hands it over casually] Fair warning though---there's a coffee stain on the back.

Figure 12: Script refinement snippets. Script 1, Script 2, and Script 3 show representative before/after edits. Red text is removed and green text is added by the iterative writing loop. Across snippets, refinement removes unnecessary dialogue, deepens the premise, and sharpens the payoff.

islands and G generations, the total writing-stage cost is:

$$\mathcal{O}\big(G \cdot K \cdot |S_k|^2 \cdot |\mathcal{C}_k|\big)\,. \tag{11}$$

Crucially, this remains lower than a globally pooled tournament over all $K|S_k|$ scripts, which would require $\mathcal{O}(G \cdot K^2|S_k|^2 \cdot |\mathcal{C}_k|)$ evaluations.

For each selected script, we generate D scene directions and refine each of the N shots over $|\mathcal{C}_{\text{render}}|$ critic iterations, yielding $\mathcal{O}(D \cdot N \cdot |\mathcal{C}_{\text{render}}|)$ render calls per script. Shot-level tournament selection adds $\mathcal{O}(D \cdot N \cdot |\mathcal{C}_{\text{render}}|^2)$ comparisons (single-elimination), and the final video-level tournament over D complete videos costs $\mathcal{O}(D \cdot |\mathcal{C}_{\text{render}}|)$ comparisons. The dominant term for the rendering stage is therefore:

$$\mathcal{O}\big(D \cdot N \cdot |\mathcal{C}_{\text{render}}|^2\big) \quad \text{render calls per script.} \tag{12}$$

In practice, the base configuration takes approximately one day to run on a single H200 GPU with an API budget of around $5, which is evidently lower than the cost of producing traditional comedy shows. Furthermore, the system is highly parallelizable. For example, we can parallelize API or model calls with respect to multiple islands and storyboards.

E Script Refinement

Figure 12 visualizes representative changes made during iterative script refinement. The writing critics remove redundant exchanges, introduce new details, and strengthen character reactions to support comedic escalation.

F STORYBOARD OUTPUT STRUCTURE

The scene director agent outputs a structured JSON object that serves as the complete production specification for the video rendering stage. Given a script and user-specified character and background assets, the agent extracts all characters and backgrounds, generates viewpoint variations for each base background (*e.g.*, turning back, left, and right), and divides the script into shots. The full schema is shown in Listing 13.

```
{
  "characters": [{
    "name": "...",            // Full name from script
    "description": "...",     // Physical and personality details
    "portrait_path": "...",   // Path to reference image, or null
    "voice_path": "...",      // Path to voice sample, or null
    "t2i_prompt": "...",      // Image gen prompt (if no portrait_path)
    "t2a_prompt": "...",      // Voice gen prompt (if no voice_path)
    "is_user_specified": true
  }],
  "backgrounds": [{
    "name": "...",            // Unique name, e.g., "Lab Front View"
    "description": "...",
    "image_path": "...",      // Path to reference image, or null
    "t2i_prompt": "...",      // Full scene prompt for base backgrounds
    "base_background": "...", // Parent background name (variations only)
    "variation_prompt": "...",// Viewpoint edit prompt (variations only)
    "is_user_specified": false
  }],
  "shots": [{
    "shot_id": "shot_01",
    "speaker": "...",         // Exact character name, or null
    "line": "...",            // Dialogue text
    "voice_intensity": 0.5,   // 0.0 (low) to 1.5 (extreme)
    "first_frame": {
      "setup_notes": "...",
      "reference": {
        "reference_characters": ["..."], // Visible characters
        "reference_backgrounds": ["..."],// Background behind speaker
        "reference_shots": ["..."],      // Prior shots for continuity
        "edit_prompt": "..."  // Static composition
      },
      "generation_prompt": null // Non-null only if no references
    },
    "video_prompt": "..."     // Camera motion and actions
  }]
}
```

Figure 13: Storyboard output structure produced by the scene director.

G SCRIPT INSPECTION AND SELECTION

After all island iterations, scripts are inspected and a final evaluation is conducted by merging all islands into a single league. To select the top scripts for rendering, we run a round-robin tournament where each script competes against all others before a specialized critic committee tailored for an academic audience. Following this, the top four scripts are selected for production.

H HUMAN EVALUATION PROTOCOL

Baseline comparison. We evaluated five methods: COMIC (ours), VGoT (Zheng et al., 2024), MovieAgent (Wu et al., 2025), Veo 3.1 (Google DeepMind, 2025), and Sora 2 (OpenAI, 2025). We conducted a baseline video evaluation with participants across the US, Europe, and Asia, yielding $110 = 22 \times 5$ responses in total. Each method was represented by four videos, without manual curation. Figure 14 illustrates the rating distributions, where we were able to obtain highly diverse opinions. Participants rated each video on a 7-point Likert scale across the following dimensions:

1. **Funniness:** "How funny did you find this video?"
   (1) Not funny at all → (2) Not very funny → (3) Slightly funny → (4) Moderately funny → (5) Quite funny → (6) Very funny → (7) Extremely funny

2. **Re-watch Intent:** "Would you like to watch more videos like this?"
   (1) Definitely not → (2) Probably not → (3) Unlikely → (4) Neutral → (5) Likely → (6) Probably yes → (7) Definitely yes
3. **Comparison to Human Comedy:** "Thinking of all the human-made sketch comedies you have ever seen, how funny is this video compared to the average human-made one?"
   (1) Much less funny → (2) Less funny → (3) Slightly less funny → (4) Comparable → (5) Slightly funnier → (6) Funnier → (7) Much funnier
4. **Script Quality:** "How would you rate the funniness of the script?"
   (1) Not funny at all → (2) Slightly funny → (3) Somewhat funny → (4) Moderately funny → (5) Quite funny → (6) Very funny → (7) Extremely funny
5. **Narrative Quality:** "How would you rate the narrative quality of the script (*e.g.*, story arc, pacing)?"
   (1) Very poor → (2) Poor → (3) Slightly poor → (4) Neutral → (5) Slightly good → (6) Good → (7) Excellent
6. **Visual Realism:** "How would you rate the visual realism of this video?"
   (1) Very poor → (2) Poor → (3) Slightly poor → (4) Neutral → (5) Slightly good → (6) Good → (7) Excellent
7. **Visual Consistency:** "How would you rate the visual consistency of this video (*e.g.*, characters, backgrounds)?"
   (1) Very poor → (2) Poor → (3) Slightly poor → (4) Neutral → (5) Slightly good → (6) Good → (7) Excellent

Ablation study. Participants completed a paired A/B comparison between COMIC and the critic-free baseline across five dimensions:

1. **Funniness:** "Which video do you find funnier?"
2. **Script Quality:** "Which video has a funnier script?"
3. **Narrative Quality:** "Which video has better narrative flow (*e.g.*, story arc, pacing)?"
4. **Visual Realism:** "Which video has higher realism?"
5. **Visual Consistency:** "Which video has higher visual consistency (*e.g.*, characters, backgrounds)?"

Each question offered three choices: *Video A*, *About the same*, or *Video B*. Neutral responses (*About the same*) were excluded from the analysis.

AI USE STATEMENT

We used generative AI tools to polish the manuscript, refine grammar, format LaTeX, and assist with coding and implementation. We did not use these tools to develop the research framework, propose or refine research hypotheses, design experiments, process the collected dataset, or interpret results. The use of generative models within COMIC for content generation and evaluation is described in the methods and model specifications. The authors take responsibility for the final manuscript, implementation, and reported results.

ETHICS STATEMENT

COMIC is intended for creative production of fictional sketch comedy. The characters in our generated videos are fictional. Nevertheless, systems that synthesize realistic audiovisual content could be misused for deepfakes, impersonation, or misleading representations. Generated content should be clearly identified as synthetic, and use of real individuals' likenesses or voices would require appropriate consent. Comedy generation can also reproduce stereotypes or produce offensive content; human review remains important before public distribution. Our intended use is fictional entertainment, and we do not endorse deceptive or harmful uses of the framework.

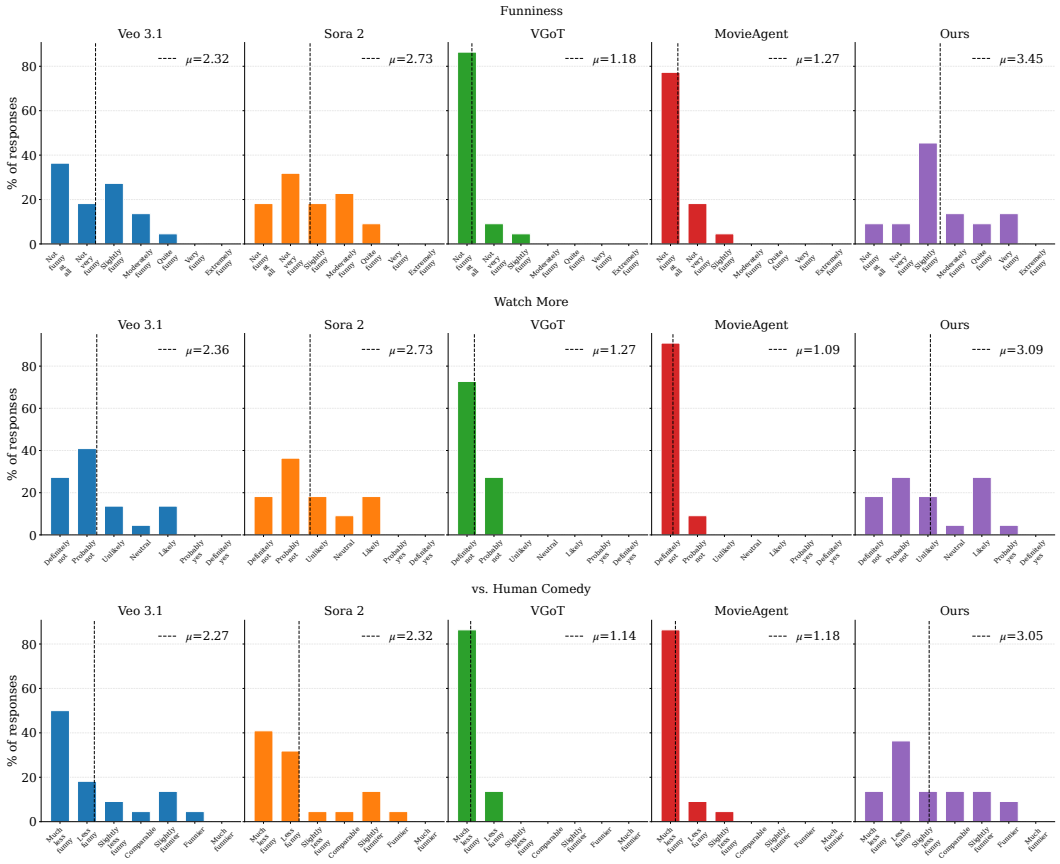


Figure 14: Distributions of human evaluation ratings for COMIC compared to baseline methods.

REPRODUCIBILITY STATEMENT

We document the method, model choices, and hyperparameter configurations to support reproduction of our experiments. Table 7 lists the models used at each stage, and Table 8 specifies the small, base, and large configurations. The appendix also describes writing-critic selection, the storyboard output structure, and the human evaluation protocol. These details represent our best effort to make the experimental setup reproducible; stochastic generation and changes to externally hosted models may affect exact outputs. We will release the full codebase upon acceptance.