

SCMAPR: Self-Correcting Multi-Agent Prompt Refinement for Complex-Scenario Text-to-Video Generation

Chengyi Yang^{1,2*}, Pengzhen Li¹, Jiayin Qi³, Aimin Zhou², Ji Wu⁴, Ji Liu^{1†}

¹HiThink Research, ²East China Normal University,
³Guangzhou University, ⁴Tsinghua University

Abstract

Text-to-Video (T2V) generation has benefited from recent advances in diffusion models, yet current systems still struggle under complex scenarios, which are generally exacerbated by the ambiguity and underspecification of text prompts. In this work, we formulate complex-scenario prompt refinement as a stage-wise multi-agent refinement process and propose SCMAPR, i.e., a scenario-aware and Self-Correcting Multi-Agent Prompt Refinement framework for T2V prompting. SCMAPR coordinates specialized agents to (i) route each prompt to a taxonomy-grounded scenario for strategy selection, (ii) synthesize scenario-aware rewriting policies and perform policy-conditioned refinement, and (iii) conduct structured semantic verification that triggers conditional revision when semantic fidelity violations are detected. To clarify what constitutes complex scenarios in T2V prompting, provide representative examples, and enable rigorous evaluation under such challenging conditions, we further introduce T2V-Complexity, which is a complex-scenario T2V benchmark consisting exclusively of complex-scenario prompts. Extensive experiments on 3 existing benchmarks and our T2V-Complexity benchmark demonstrate that SCMAPR consistently improves text-video alignment and overall generation quality under complex scenarios, achieving up to 2.67% and 3.28 gains in average score on VBench and EvalCrafter, and up to 0.028 improvement on T2V-CompBench over 3 state-of-the-art baselines. The codes of SCMAPR are publicly available at <https://github.com/HiThink-Research/SCMAPR>.

1 Introduction

The rapid advancement of diffusion models (Peebles and Xie, 2023; Rombach et al., 2022) has rev-

olutionized Artificial Intelligence Generated Content (AIGC), with applications ranging from image and video generation to 3D content creation (Guo et al., 2025; Huang et al., 2025b; Lin et al., 2025; Zhang et al., 2025a), speech and audio synthesis (Luo et al., 2023; Liu et al., 2024b; Oh et al., 2024), and controllable editing (Lee et al., 2025; He et al., 2025; Wang et al., 2025d), driving new opportunities in entertainment, education, design, and human-computer interaction. As an AIGC modality, Text-to-Video (T2V) generation requires visually realistic frames together with temporal coherence, motion dynamics, and adherence to physical and causal constraints, thereby posing a substantially challenging yet impactful open research problem.

Existing studies (Hao et al., 2023; Zhan et al., 2024b,a) demonstrate that optimizing prompts with Large Language Models (LLMs) leads to high-quality and user-aligned generations in diffusion-based content creation. Such improvements are particularly evident when user inputs are concise and underspecified. In practice, high-quality prompts specify characters and scenes precisely, follow effective expression patterns, and incorporate domain-specific terminology for stylistic control (Parsons, 2022; Witteveen and Andrews, 2022; Brade et al., 2023; Zhan et al., 2024a).

However, not all T2V generation tasks can be materially improved by merely expanding the input text. Complex-scenario T2V generation tasks differ from common T2V tasks in the target video scenario. For complex-scenario T2V generation tasks, the target video scenario is dominated by one or multiple specific challenging sources for the current T2V diffusion model to realize with temporal coherence. To make this notion explicit, we introduce a ten-category taxonomy of complex scenarios and use the dominant category as a routing signal for policy generation and prompt refinement. Please see formal definitions and representative examples in Appendix A.

*Work done during his internship at HiThink Research under supervision of Ji Liu.

†Corresponding author: jiliuwork@gmail.com

Meanwhile, existing prompt refinement approaches are generally designed for Text-to-Image (T2I) tasks, which typically involve employing retrieval-augmented generation to expand prompts (Sun et al., 2024), tailoring prompts based on user preferences (Zhan et al., 2024b,a), and enhancing prompts through entity-specific descriptions (Ozaki et al., 2025). While effective for single-image synthesis, these strategies are insufficient for T2V generation. Unlike T2I, T2V tasks are required to simultaneously guarantee temporal consistency, ensure motion coherence, capture causal dependencies, and comply with physical laws across frames. Although a few prompt refinement approaches exist for T2V, e.g., Retrieval-Augmented Prompt Optimization (RAPO) (Gao et al., 2025), they may focus on inter-object relations while overlooking video-specific challenges, e.g., abstract semantics and temporal consistency, in complex scenarios. In addition, existing T2V benchmarks are not designed for complex scenarios. They typically do not fully cover complex-scenario categories and exhibit severe category imbalance, which hinders rigorous and systematic evaluation under complex scenarios.

In this paper, we formulate the prompt refinement process for complex-scenario T2V generation as a *stage-wise multi-agent refinement process*. In this process, specialized agents collaboratively identify the dominant source of difficulty via scenario tagging. Then, they plan scenario-appropriate refinement strategies and verify semantic fidelity through structured analysis. Motivated by this perspective, we develop a Self-Correcting Multi-Agent Prompt Refinement framework (SCMAPR) with a pipeline of five stages, i.e., (1) taxonomy-based scenario routing, (2) scenario-aware policy synthesis, (3) policy-conditioned prompt refinement, (4) structured semantic verification, and (5) conditional revision according to validation results. To clarify complex scenarios in T2V prompts with representative examples, we introduce the benchmark T2V-Complexity, consisting exclusively of prompts balanced across complex-scenario categories. We note that multi-agent prompt refinement has been explored for complex T2I generation (Li et al., 2025). In contrast, our method targets T2V and introduces verification-driven self-correction rather than pure information extraction.

Overall, SCSMAPR is designed to enable scenario-aware and self-correcting prompt refinement for complex-scenario T2V generation.

SCMAPR exploits structured and verification-driven signals to guide conditional revision, thereby improving semantic fidelity under challenging prompts. Together with the proposed T2V-Complexity benchmark, SCSMAPR supports systematic study and evaluation of complex-scenario T2V prompting. The main contributions are summarized as follows:

- We introduce a scenario-aware refinement pipeline for complex-scenario T2V prompt refinement, including taxonomy-grounded routing, prompt-specific policy synthesis, and policy-conditioned rewriting (Stages I-III).
- We propose a verification-driven self-correction design that performs atom-level semantic verification and conditional prompt revision to improve user-intent preservation and semantic fidelity (Stages IV-V).
- We introduce T2V-Complexity, a taxonomy-grounded benchmark consisting exclusively of complex-scenario prompts and balanced across ten complex-scenario categories, enabling rigorous evaluation of T2V generation under diverse complex scenarios.
- We conduct extensive experimentation on 3 existing benchmarks and T2V-Complexity, demonstrating that SCSMAPR consistently improves semantic fidelity and overall video generation quality for T2V generation.

2 Related Work

In this section, we present existing works on T2V generation and prompt refinement approaches.

2.1 Text-to-Video Generation

Recent advancements in diffusion transformers and large-scale generative models (Rombach et al., 2022; Peebles and Xie, 2023) have significantly promoted T2V generation (Singer et al., 2023; Chen et al., 2024a). Prior works have explored a variety of directions, including scalable architectures such as expert transformers and linear-complexity attention modules (Yang et al., 2025; Wang et al., 2025b), training-free and plug-and-play inference techniques for improving motion dynamics and spatial fidelity (Bu et al., 2025; Zhang et al., 2025b; Jagpal et al., 2025), as well as structured captions, instance-aware modeling and Low-Rank Adaptation (LoRA)-based customization for controlling

entity appearance and interactions (Feng et al., 2025; Fan et al., 2025; Huang et al., 2025a). Beyond visual quality, LLM-guided reasoning and external knowledge retrieval have also been introduced to enhance physical plausibility and factual correctness (Xue et al., 2025; Yuan et al., 2025). Despite these advances, most existing T2V approaches implicitly assume relatively common scenarios with well-specified and explicit prompts. As a result, T2V generation under complex scenarios, including those involving abstract semantics, intricate multi-entity interactions, and long-range temporal dependencies, remains challenging.

2.2 Prompt Refinement

Prompt refinement aims to transform user inputs into formulated prompts that align with the preferences and capabilities of diffusion models. Early studies focus on inferring user preferences or rewriting patterns to guide prompt refinement. Representative approaches, such as PRIP (Zhan et al., 2024b) and CAPR (Zhan et al., 2024a), adapt prompts based on inferred user capabilities or configurable features, but rely heavily on user interaction data, system logs, or large-scale feedback, which limits their applicability in settings without explicit user supervision. More recent approaches incorporate Retrieval-Augmented Generation (RAG) to enrich prompts by retrieving semantically relevant descriptions. These methods either maintain external prompt repositories (Sun et al., 2024) or construct relation graphs from training data to retrieve semantically related terms (Gao et al., 2025), which are subsequently processed and incorporated into the user input to produce an expanded reformulation. While effective for improving descriptive richness, such relevance-driven retrieval and augmentation strategies primarily expand prompts based on surface similarity, without explicitly reasoning about the underlying sources of difficulty or specifying scenario-specific refinement guidelines.

3 Self-Correcting Multi-Agent Prompt Refinement

In this section, we first present the stage-wise agent-based framework of SCMAPR. Then, we present the details of each stage.

3.1 Stage-wise Agent-Based Architecture

T2V prompts impose heterogeneous reasoning demands that go beyond surface attributes such as

length or visual richness. In addition, their difficulty arises in complex scenarios where user intent is under-specified, abstract, or internally entangled. Complex scenarios pose challenges to constructing temporally coherent and semantically faithful video descriptions. A key obstacle is that complex-scenario prompt refinement typically involves multiple coupled constraints. It requires clarifying intent, organizing entities and events into a coherent spatio-temporal structure, and preserving semantic fidelity throughout rewriting. When these requirements are handled implicitly within a single rewrite, the prompt refinement tends to become brittle. As a result, semantic omissions and contradictions may be introduced relative to the user input.

Motivated by this observation, we construct a stage-wise multi-agent pipeline within SCMAPR, which externalizes the prompt refinement into 5 explicit stages with intermediate representations and verification-driven feedback. The stage-wise multi-agent pipeline incorporates conditional self-correction to achieve excellent T2V generation, in which specialized agents communicate through explicit intermediate representations. As illustrated in Figure 1, SCMAPR comprises six functional agents organized into two interacting groups. First, the *refinement group* includes a Scenario Router, a Policy Generator, and a Prompt Refiner, which jointly perform scenario-aware strategy selection and policy-conditioned prompt rewriting. Second, the *verification group* consists of a Semantic Atomizer, an Entailment Validator and a Content Reviser, which collectively enforce semantic fidelity through atom-level verification. In the common case, refinement proceeds forward from routing and policy generation to rewriting and verification. When semantic fidelity violations are detected, structured feedback from the Entailment Validator selectively triggers revision by Content Reviser, while atomic constraints remain fixed.

3.2 Stage I: Scenario Tagging for Strategy Routing

SCMAPR exploits a predefined routing tag set comprising 11 scenario labels of two types, i.e., a non-difficult tag and ten complex-scenario tags: Abstract Descriptions, Complex Spatial Relations, Multi-Element Scenes, Fine-Grained Appearance, Temporal Consistency, Stylistic Hybrids, Causality & Physics, Camera Motion, Object Interaction, and Scene Transitions (see Appendix A for formal definitions and representative examples).

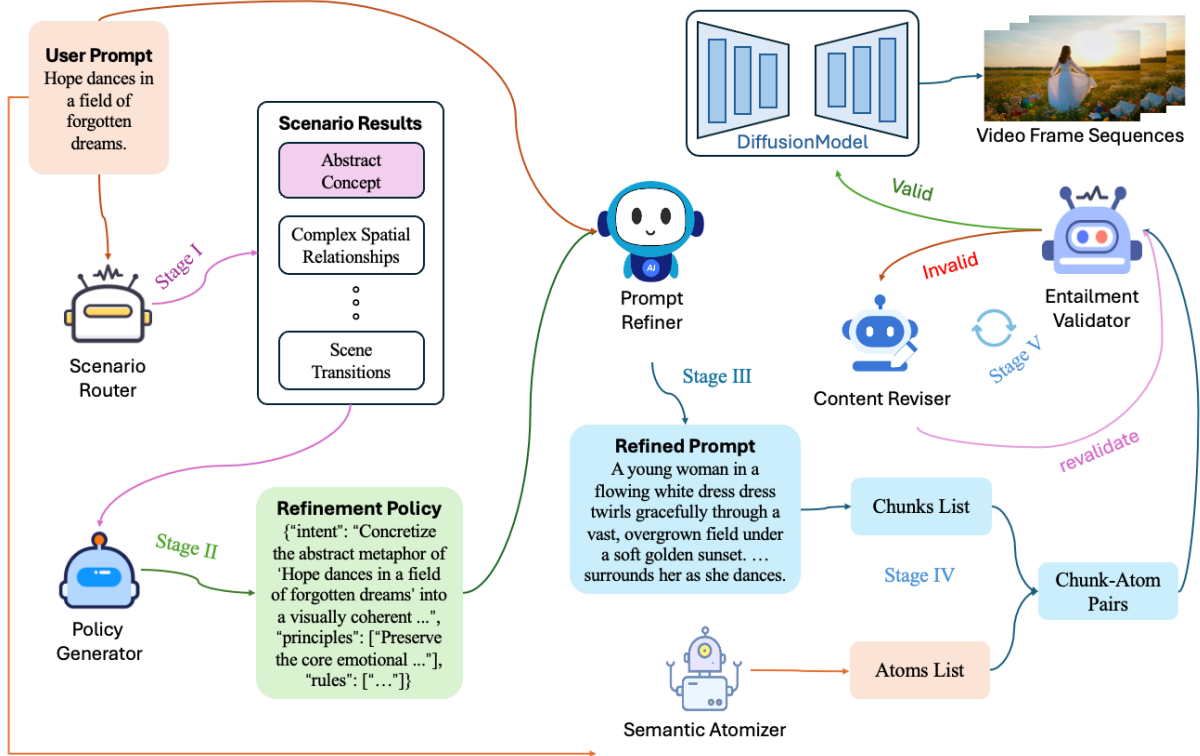


Figure 1: **Self-Correcting Multi-Agent Prompt Refinement Framework (SCMAPR)**. SCMAPR organizes prompt refinement as a stage-wise multi-agent collaboration involving six specialized agents. The framework proceeds through five functional stages: (I) *Scenario Routing*, where Scenario Router assigns a scenario tag to the input prompt. (II) *Policy Synthesis*, where a Policy Generator generates a scenario-conditioned rewriting policy. (III) *Policy-Conditioned Refinement*, where a Prompt Refiner rewrites the prompt. (IV) *Semantic Verification*, where Atomizer and Validator collaboratively verify semantic fidelity through atomic extraction and entailment judgment. (V) *Conditional Revision*, where verification feedback conditionally triggers targeted revision, enabling self-correcting refinement.

The non-difficult tag serves as a conservative fallback for ordinary prompts that do not exhibit a dominant complex constraint, e.g., “An airplane.” in VBench (Huang et al., 2024), while the ten complex-scenario tags capture dominant sources of difficulty in complex-scenario T2V prompting and act as routing signals for downstream policy synthesis and prompt refinement. Formal definitions and representative examples of the ten complex-scenario categories are provided in Appendix A.

Given user input P_{user} , Scenario Router classifies the input and attributes a corresponding scenario tag \hat{y} based on an LLM (see prompt details in Appendix E.1). The scenario tag is represented with a one-shot setting selected from the predefined scenario label set. Then, the routing signal, i.e., Scenario Tag \hat{y} , is passed to the Policy Generator for scenario-aware policy synthesis.

3.3 Stage II: Policy Synthesis for Scenario-Aware Refinement

SCMAPR employs the Policy Generator to synthesize a prompt-specific rewriting policy. When re-

ceiving user input and scenario tag, i.e., $(P_{\text{user}}, \hat{y})$, from Scenario Router, the Policy Generator outputs a scenario-conditioned rewriting policy π , which provides explicit guidance on how refinement should be performed under the routed scenario, exploiting an LLM (see prompt details in Appendix E.2).

Concretely, the generated policy π specifies three aspects. First, it identifies which implicit constraints should be made explicit (e.g., entities, relations, actions, temporal stages). Second, it indicates which scenario-relevant modeling competencies should be emphasized (e.g., spatial layout, temporal coherence, physical plausibility, camera motion). Third, it enforces a conservative fidelity principle that preserves the user-intended meaning and avoids introducing unsupported content.

Scenario Tag \hat{y} is provided as a conditioning context for policy synthesis, while preserving flexibility across prompts. This design enables refinement strategies to be dynamically synthesized rather than pre-specified.

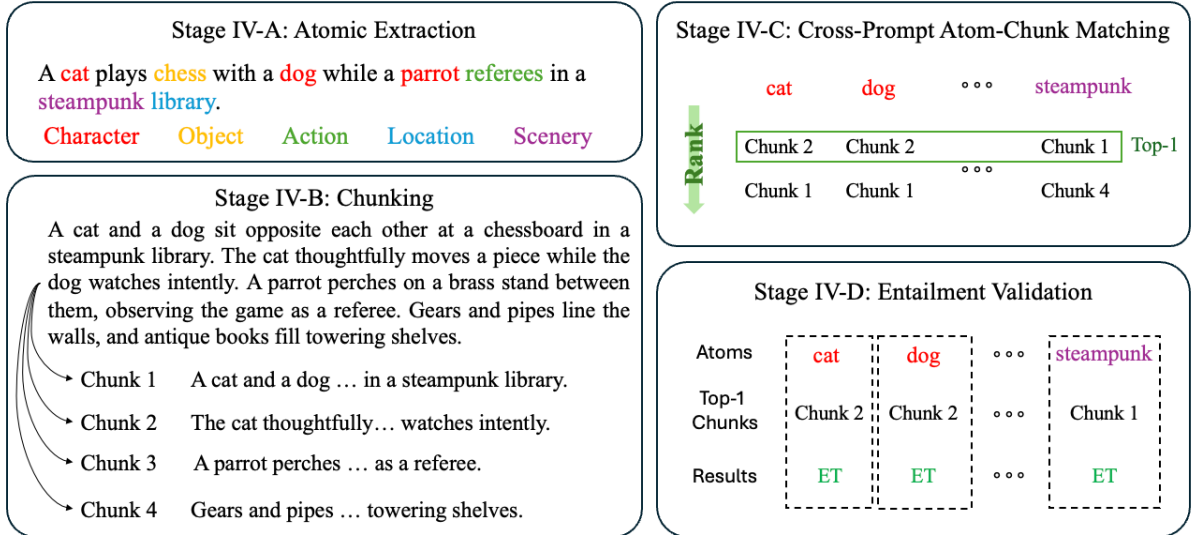


Figure 2: **Illustration of the Semantic Verification Stage in SCMAPR.** Given a user input and the corresponding refined prompt, semantic verification is performed in four steps. (1) *Atomic Extraction* decomposes the user input into atom elements. (2) *Chunking* segments the refined prompt into semantically coherent evidence units. (3) *Atom-Chunk Matching* retrieves the most relevant evidence chunk for each atom. (4) *Entailment Validation* assesses atom-level semantic relations between atoms and evidence chunks. Through this design, semantic missing and contradictions in the refined prompt can be detected and subsequently used to trigger downstream revision.

3.4 Stage III: Policy-Conditioned Prompt Refinement

Conditioned on the generated policy π and the original user input P_{user} , the Prompt Refiner generates a refined prompt P_{rew} based on an LLM (see prompt details in Appendix E.3). The Refiner rewrites P_{user} into a clearer and more model-friendly description by elaborating underspecified details, e.g., characters, objects, actions, and settings, while preserving the original intent, tone, and stated content. Guided by π , the refiner avoids introducing unsupported elements and produces the refined prompt as a small number of concise sentences, without additional explanations. By decoupling policy synthesis from prompt rewriting, SCMAPR encourages faithful execution of explicit constraints and mitigates semantic deviation during refinement.

3.5 Stage IV: Semantic Verification via Atomization and Entailment Validation

To ensure semantic fidelity, SCMAPR performs structured verification on the refined prompt. This stage is realized by Semantic Atomizer and Entailment Validator, which interact through explicit intermediate representations to assess consistency between P_{user} and P_{rew} . Details are available in Figure 2.

Stage IV-A: Atomic Extraction as Fixed Verification Targets. Given the user input P_{user} , Seman-

tic Atomizer extracts a field-wise atom dictionary $\mathcal{D}_{\mathcal{A}} = \text{Atomizer}(P_{\text{user}})$ based on an LLM (see prompt details in Appendix E.4). This dictionary contains five fields, namely characters, objects, actions, locations and scenery. For instance, given “A cat plays chess with a dog while a parrot referees in a steampunk library” as user input, Semantic Atomizer produces $\mathcal{D}_{\mathcal{A}} = \{ \text{characters: [cat, dog, parrot]}, \text{objects: [chess]}, \text{actions: [plays, referees]}, \text{locations: [library]}, \text{scenery: [steampunk]} \}$. We treat $\mathcal{D}_{\mathcal{A}}$ as fixed verification targets throughout subsequent stages. For subsequent matching, we flatten the dictionary into a single list $\mathcal{A} = \text{Flatten}(\mathcal{D}_{\mathcal{A}}) = \{a_1, a_2, \dots, a_n\}$, where a_i is an extracted atom element (e.g., $a_1 = \text{"cat"}$) and n represents the number of atom elements in P_{user} .

Stage IV-B: Sentence-Level Chunking. The refined prompt P_{rew} is segmented into non-overlapping chunks $\mathcal{C} = \text{Chunk}(P_{\text{rew}}) = \{c_1, \dots, c_m\}$ with m referring to the number of chunks in P_{rew} . By default, each sentence forms a chunk. Short sentences are merged with subsequent ones until a length threshold is reached.

Stage IV-C: Evidence Selection via Atom-Chunk Matching. Given atom elements $\{a_i\}_{i=1}^n$ and chunks $\{c_j\}_{j=1}^m$, we embed atom elements and chunks into a shared semantic space using a lightweight embedding model $f_{\mathcal{E}}(\cdot)$ and compute

Table 1: Quantitative results (%) on VBench with LaVie and Wan as T2V backbones. Bold values denote the best performance for each metric under each backbone.

Method	Average Score	Aesthetic Quality	Background Consistency	Imaging Quality	Motion Smoothness	Subject Consistency	Temporal Flickering
LaVie	81.89	53.52	95.01	62.83	95.15	90.96	93.85
LaVie + Open-Sora	81.95	53.67	94.82	62.93	95.28	91.24	93.78
LaVie + RAPO	82.80	54.38	95.76	63.06	96.03	93.21	94.35
LaVie + SCMAPR	84.56	56.53	97.43	63.87	97.18	95.95	96.39
Wan	86.19	63.75	95.41	69.78	97.62	95.12	95.43
Wan + Open-Sora	86.27	63.82	95.22	70.05	97.76	95.28	95.52
Wan + RAPO	87.43	64.12	97.35	70.99	98.72	96.18	97.23
Wan + SCMAPR	88.21	65.19	98.04	71.67	98.98	97.20	98.19

Table 2: Quantitative comparisons on EvalCrafter. SCMAPR consistently achieves better results, demonstrating a clear lead on this benchmark.

Method	Average	Motion Quality	Text-Video Alignment	Visual Quality	Temporal Consistency
LaVie	62.12	53.19	69.60	64.81	60.87
LaVie + Open-Sora	62.78	53.07	71.38	65.26	61.41
LaVie + RAPO	63.91	53.34	74.38	66.62	61.29
LaVie + SCMAPR	65.18	53.87	75.42	68.84	62.56
Wan	63.46	53.79	73.05	65.17	61.85
Wan + Open-Sora	63.95	53.93	73.84	65.60	62.43
Wan + RAPO	64.54	54.26	75.13	66.08	62.71
Wan + SCMAPR	66.74	54.85	76.94	70.23	64.93

cosine similarity $s_{ij} = \cos(f_{\mathcal{E}}(a_i), f_{\mathcal{E}}(c_j))$. For atom element a_i , corresponding evidence chunk e_i is selected by maximizing the cosine similarity, i.e., $e_i = \arg \max_{j \in \{1, \dots, m\}} s_{ij}$.

Stage IV-D: Atom-Level Entailment Validation.

For each atom element $a_i \in \mathcal{A}$, we pair it with the selected evidence chunk $e_i \in \mathcal{C}$ and invoke Entailment Validator to assess whether the semantics expressed by a_i are supported by e_i with an LLM (see prompt details in Appendix E.5). Formally, the validator generates a ternary decision $v_i = \text{Validator}(a_i, e_i)$, where $v_i \in \{\text{ET}, \text{MS}, \text{CT}\}$ denotes the entailment label for atom a_i with respect to chunk e_i , i.e., *Entailment* (ET), *MisSing* (MS), and *ConTradiction* (CT). ET indicates that the semantics of a_i are well supported by e_i . MS refers to the case, in which e_i provides little support on the semantic expression of a_i . CT indicates that e_i has opposite semantic meanings compared to a_i .

3.6 Stage V: Conditional Revision

To monitor verification quality, we compute the coverage rate p_{ET} and contradiction rate p_{CT} as diagnostic statistics after atom-level entailment judgments. Specifically, they measure the fractions of atom-chunk pairs labeled as entailment and contra-

diction, respectively:

$$p_{\text{ET}} = \frac{1}{|\mathcal{A}|} \sum_{a_i \in \mathcal{A}} \mathbb{I}[\text{Validator}(a_i, e_i) = \text{ET}], \quad (1)$$

$$p_{\text{CT}} = \frac{1}{|\mathcal{A}|} \sum_{a_i \in \mathcal{A}} \mathbb{I}[\text{Validator}(a_i, e_i) = \text{CT}]. \quad (2)$$

In our framework, p_{ET} and p_{CT} are used for diagnostic monitoring rather than as soft indicators for revision decisions. We enforce a strict acceptance criterion that the refinement is accepted only if $p_{\text{ET}} = 1$ (100%) and $p_{\text{CT}} = 0$. Otherwise, Entailment Validator returns structured feedback pinpointing missing or contradicted atoms, which triggers Content Reviser to revise the prompt and produce an updated P_{rew} . The procedure repeats until the acceptance criterion is met or a maximum number of revision rounds is reached. In many cases, the criterion is satisfied immediately, and the framework completes in a single forward pass.

4 T2V-Complexity Benchmark

To enable rigorous evaluation of T2V generation under taxonomy-defined complex scenarios, we introduce *T2V-Complexity*. This benchmark contains 1000 user-style prompts, with 100 prompts for each of the ten complex-scenario categories in

Table 3: Quantitative comparisons on T2V-CompBench. SCMAPR achieves the highest average score.

Method	Average	Consistent Attribute	Dynamic Attribute	Action Binding	Motion Binding
LaVie	0.388	0.620	0.232	0.483	0.215
LaVie + Open-Sora	0.361	0.532	0.214	0.470	0.226
LaVie + RAPO	0.460	0.692	0.267	0.635	0.243
LaVie + SCMAPR	0.476	0.704	0.273	0.640	0.285
Wan	0.454	0.694	0.263	0.591	0.269
Wan + Open-Sora	0.446	0.672	0.258	0.583	0.274
Wan + RAPO	0.495	0.721	0.279	0.672	0.309
Wan + SCMAPR	0.523	0.756	0.297	0.691	0.346

Table 4: Quantitative results (%) on T2V-Complexity using VBench metrics with Wan as the T2V backbone.

Method	Average Score	Aesthetic Quality	Background Consistency	Imaging Quality	Motion Smoothness	Subject Consistency	Temporal Flickering
Wan	82.95	56.95	94.83	63.80	96.27	91.28	94.59
Wan + SCMAPR	85.69	59.32	96.24	67.87	98.87	93.94	97.92

Table 5: Statistics of complex-scenario prompts across different T2V benchmarks. Here, CS is the abbreviation for complex-scenario.

T2V Benchmark	Prompts	CS Prompts	CS Proportion	CS Categories
VBench	946	311	32.9%	9
EvalCrafter	700	422	60.3%	9
T2V-CompBench	1400	972	69.4%	10
T2V-Complexity	1000	1000	100.0%	10

Table 6: Ablation studies of different components in SCMAPR on VBench.

Method	Average Score
SCMAPR	88.21%
w/o Scenario Routing	86.49%
w/o Policy Generation	87.75%
w/o Verification & Self-Correction	87.63%

our taxonomy. Each prompt corresponds to a target generation task in its designated category while accompanied by expected failure modes, which enables interpretable and fine-grained analysis.

In contrast to prior benchmarks that contain a mixture of ordinary and complex prompts, T2V-Complexity is constructed to focus exclusively on complex-scenario prompts. Table 5 shows that VBench (946 prompts), EvalCrafter (700 prompts) and T2V-CompBench (1400 prompts) contain 32.9%, 60.3%, and 69.4% complex-scenario prompts, respectively. Crucially, T2V-Complexity comprises 1000 prompts that are all complex-scenario prompts and covers 10 complex-scenario categories with a balanced design. This design aligns with our motivation that existing benchmarks may provide limited coverage of certain complex categories and exhibit severe scenario imbalance (also see Appendix D.1), which poses challenges for fair and effective evaluation.

5 Experiments

In this section, we present the experimental settings with three benchmarks. Then, we demonstrate the main experimental results. Finally, we show an ablation study.

5.1 Experimental Setup

Benchmarks: We conduct evaluations on three existing State-Of-The-Art (SOTA) benchmarks, i.e., VBench (Huang et al., 2024), EvalCrafter (Liu et al., 2024c), and T2V-CompBench (Sun et al., 2025), to evaluate the quality of T2V generation. In addition, we further evaluate SCMAPR on our T2V-Complexity, which is designed for complex scenarios. All table metrics are averaged over prompts.

Baselines: We compare three SOTA representative prompt refinement approaches in the field of T2V, including the direct prompting (using user input), prompt refiner from Open-Sora (Zheng et al., 2024), and RAPO (Gao et al., 2025).

Implementation: In our experiments, we adopt DeepSeek-V3.2 (DeepSeek-AI et al., 2025) to construct multiple agents. In addition, we employ BGE-M3 (Chen et al., 2024b) as the embedding model in atom-chunk matching. For video generation, we adopt Wan2.2 (Wan) (Wang et al., 2025a) and LaVie (Wang et al., 2025c) as T2V backbones.

5.2 Main Results

As shown in Tables 1, 2 and 3, we report the quantitative results on three benchmarks, using LaVie and Wan as T2V backbones. Under both backbones, SCMAPR consistently achieves the best overall performance, demonstrating its effectiveness for T2V generation. On VBench, SCMAPR achieves the

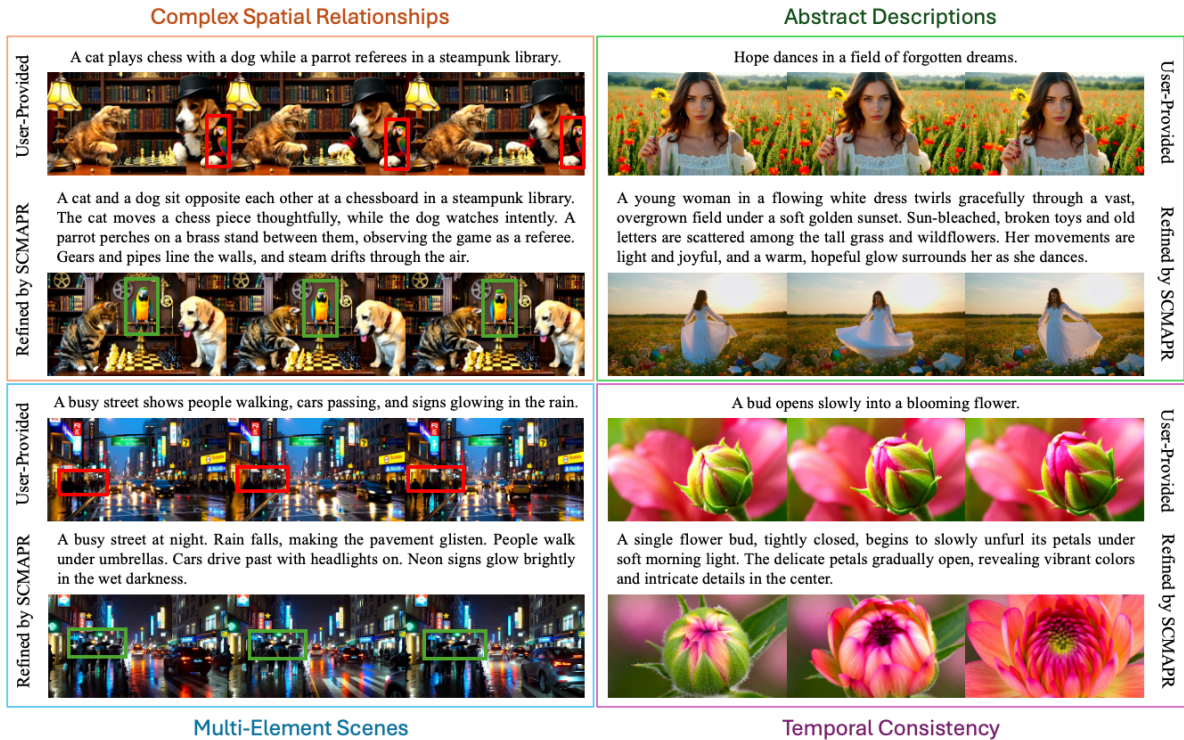


Figure 3: Comparisons of videos generated using Wan (Wang et al., 2025a) conditioned on user input and refined prompts from SCMAPR.

highest average score of 84.56% with LaVie and 88.21% with Wan. Compared with direct prompting, Open-Sora and RAPO, SCMAPR improves the average score by 2.67%/2.02%, 2.61%/1.94%, and 1.76%/0.78% under LaVie/Wan, respectively. We can observe similar trends on EvalCrafter, where SCMAPR achieves the highest average score of 65.18 with LaVie and 66.74 with Wan. Compared with direct prompting, Open-Sora and RAPO, SCMAPR improves the average score by 3.06/3.28, 2.40/2.79 and 1.27/2.20 under LaVie/Wan, respectively. Notably, it demonstrates consistent improvements in T2V alignment and temporal consistency, underscoring its strength in semantic fidelity and dynamic coherence. When it comes to T2V-CompBench with an emphasis on compositional reasoning, SCMAPR achieves the best results in consistent attribute binding (up to 0.035 higher), dynamic attribute binding (up to 0.018 higher), action binding (up to 0.019 higher), and motion binding (up to 0.042 higher), demonstrating excellent performance in fine-grained attribute and motion modeling. In particular, SCMAPR significantly surpasses RAPO, Open-Sora, and direct prompting by 0.016/0.028, 0.115/0.077, and 0.088/0.069 on T2V-CompBench in terms of average score under LaVie/Wan, respectively.

Results on T2V-Complexity. We further eval-

uate SCMAPR on T2V-Complexity, which consists exclusively of complex-scenario prompts. As shown in Table 4, adopting the VBench metrics, SCMAPR improves the average score from 82.95% to 85.69%, yielding a 2.74% gain over direct prompting. Moreover, SCMAPR achieves consistent improvements across all six metrics, including aesthetic quality, background consistency, imaging quality, motion smoothness, subject consistency, and temporal flickering. These results further verify that SCMAPR remains effective when evaluation is restricted to exclusively complex-scenario prompts, leading to broadly improved visual quality and temporal stability under challenging scenarios.

To provide an intuitive demonstration on the advantages of SCMAPR, Figure 3 highlights the effectiveness of SCMAPR across four categories of complex scenarios: complex spatial relations, abstract descriptions, multi-entity scenes and temporal consistency. In the case of complex spatial relations, SCMAPR achieves two notable improvements over videos generated directly from user input. (1) The parrot is correctly placed at the center (green box) rather than at the edge (red box). (2) The library is rendered with steampunk elements instead of being depicted as a regular library. For abstract descriptions, SCMAPR produces richer actions and vivid visual dynamics. In multi-entity scenes, SCMAPR

meticulously introduces umbrellas for pedestrians in rainy weather (green box). Finally, under temporal consistency, while direct prompting fails to present a fully blossomed flower, SCMAPR generates the complete blooming process, resulting in a fully blossomed flower video. Beyond quantitative gains, SCMAPR demonstrates effectiveness in reducing hallucinations, especially under abstract or unspecified user input. SCMAPR turns vague intentions into grounded visual constraints and coherent descriptions, which can reduce spurious content (e.g., unintended faces or irrelevant objects) and improve concept faithfulness (see details in Appendix C).

5.3 Ablation Study

We conduct ablation experiments on key components of the agentic refinement framework on VBench. Table 6 shows that the full SCMAPR achieves the best performance. Removing scenario routing leads to the largest degradation (1.72%), indicating the necessity of category-specific rewriting policies. Disabling policy generation yields a smaller drop (0.46%), suggesting that prompt-specific policy synthesis provides additional gains beyond routing alone. Finally, removing the verification and self-correction also noticeably degrades performance (0.58%), verifying that atom-level verification and conditional correction contribute to maintaining semantic fidelity.

6 Conclusion

In this paper, we propose *Self-Correcting Multi-Agent Prompt Refinement* (SCMAPR), i.e., a structured framework for refining text input under complex T2V scenarios. SCMAPR formulates prompt refinement as a stage-wise multi-agent framework, in which six specialized agents collaboratively operate across five functional stages, including scenario routing, policy synthesis, prompt refinement, semantic verification and conditional revision. By combining taxonomy-grounded routing with atom-level entailment-based verification, SCMAPR enables targeted prompt refinement while preserving complete user intent through conditional self-correction. Extensive experiments on VBench, EvalCrafter and T2V-CompBench demonstrate that SCMAPR consistently improves text-video alignment and overall generation quality under complex scenarios, achieving up to 2.67% and 3.28 gains in average score on VBench and EvalCrafter, re-

spectively, and up to 0.028 improvement on T2V-CompBench compared with state-of-the-art baseline approaches.

Limitations

SCMAPR is a training-free and model-agnostic framework that improves text-to-video generation under complex scenarios via stage-wise multi-agent collaboration. Nevertheless, several limitations remain. First, the multi-agent design introduces additional inference overhead, especially in the semantic verification stage involving atomic extraction and entailment judgment. Second, the effectiveness of scenario routing and semantic verification depends on the reasoning capability of the instruction-tuned LLM, which may lack adapted fine-tuning (Liu et al., 2024a) or reinforcement learning (Zhu et al., 2025b,a). Third, SCMAPR adopts a predefined set of scenario tags as lightweight routing signals, which capture major orthogonal sources of difficulty in T2V generation but do not preclude the incorporation of additional categories as new challenges arise. Finally, in this work, we only focus on text inputs, and extending the framework to multimodal inputs requires further adaptation (Liu et al., 2025).

Ethical Considerations

This work aims to improve prompt refinement for text-to-video generation. As a training-free framework, SCMAPR does not introduce new generative capabilities beyond the T2V model and LLM. Nevertheless, like other prompt-based systems, it may pose potential risks of misuse if deployed without appropriate safeguards. T2V-Complexity has been manually inspected to ensure that each benchmark item does not contain any information that names or uniquely identifies individual people. In practice, responsible deployment should rely on the safety policies and content filtering mechanisms of the base models, together with standard safeguards such as prompt moderation and output filtering.

Acknowledgements

Part of the work is supported by National Natural Science Foundation of China (No. 72293583), Major Project for Tackling Key Problems in Philosophy and Social Science Research of the Ministry of Education (grant number: 24JZD040), and Consulting Project of the Chinese Academy of Engineering (No. 2025-XZ-08).

References

- Stephen Brade, Bryan Wang, Maurício Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 96:1–96:14.
- Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. 2025. Bytheway: Boost your text-to-video generation model to higher quality in a training-free way. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008.
- Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024a. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7310–7320.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics (ACL Findings)*, pages 2318–2335.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, and 263 other authors. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.
- Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Zhenheng Yang, Chaoyou Fu, Xiang Li, Jian Yang, and Ying Tai. 2025. Instancecap: Improving text-to-video generation via instance-aware structured caption. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 28974–28983.
- Weixi Feng, Chao Liu, Sifei Liu, William Yang Wang, Arash Vahdat, and Weili Nie. 2025. Blobgen-vid: Compositional text-to-video generation with blob video representations. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12989–12998.
- Bingjie Gao, Xinyu Gao, Xiaoxue Wu, Yujie Zhou, Yu Qiao, Li Niu, Xinyuan Chen, and Yaohui Wang. 2025. The devil is in the prompts: Retrieval-augmented prompt optimization for text-to-video generation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3173–3183.
- Hao Guo, Xiaoshui Huang, Jiacheng Hao, Yunpeng Bai, Hongping Gan, and Yilei Shi. 2025. Brepdiff: Lightweight generation of complex b-rep with 3d GAT diffusion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 26587–26596.
- Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing prompts for text-to-image generation. In *Conf. on Neural Information Processing Systems (NeurIPS)*.
- Kai He, Chin-Hsuan Wu, and Igor Gilitschenski. 2025. CTRL-D: controllable dynamic 3d scene editing with personalized 2d diffusion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 26630–26640.
- Chi-Pin Huang, Yen-Siang Wu, Hung-Kai Chung, Kai-Po Chang, Fu-En Yang, and Yu-Chiang Frank Wang. 2025a. Videomage: Multi-subject and motion customization of text-to-video diffusion models. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 17603–17612.
- Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xi-hui Liu, Yan-Pei Cao, and Lu Sheng. 2025b. MIDI: multi-instance diffusion for single image to 3d scene generation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 23646–23657.
- Ziqi Huang, Yanan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yao-hui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818.
- Diljeet Jagpal, Xi Chen, and Vinay P. Namboodiri. 2025. EIDT-V: exploiting intersections in diffusion trajectories for model-agnostic, zero-shot, training-free text-to-video generation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 18219–18228.
- Che Hyun Lee, Heeseung Kim, Jiheum Yeom, and Sungro Yoon. 2025. Editext: Controllable coarse-to-fine text editing with diffusion language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 22798–22815.
- Mingcheng Li, Xiaolu Hou, Ziyang Liu, Dingkan Yang, Ziyun Qian, Jiawei Chen, Jinjie Wei, Yue Jiang, Qingyao Xu, and Lihua Zhang. 2025. MCCD: Multi-agent collaboration-based compositional diffusion for complex text-to-image generation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Jiantao Lin, Xin Yang, Meixi Chen, Yingjie Xu, Dongyu Yan, Leyi Wu, Xinli Xu, Lie Xu, Shunsi Zhang, and Ying-Cong Chen. 2025. Kiss3dgen: Repurposing image diffusion models for 3d asset generation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5870–5880.
- Che Liu, Yingji Zhang, Dong Zhang, Weijie Zhang, Chenggong Gong, Yu Lu, Shilin Zhou, Ziliang Gan, Ziao Wang, Haipang Wu, and 1 others. 2025. Nexus-o: An omni-perceptive and-interactive model for language, audio, and vision. In *ACM Int. Conf. on Multimedia (MM)*, pages 10787–10796.

- Ji Liu, Jiayang Ren, Ruoming Jin, Zijie Zhang, Yang Zhou, Patrick Valduriez, and Dejing Dou. 2024a. Fisher information-based efficient curriculum federated learning with large language models. In *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10497–10523.
- Weizhi Liu, Yue Li, Dongdong Lin, Hui Tian, and Haizhou Li. 2024b. GROOT: generating robust watermark for diffusion-model-based audio synthesis. In *ACM Int. Conf. on Multimedia (MM)*, pages 3294–3302.
- Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tiejong Zeng, Raymond Chan, and Ying Shan. 2024c. EvalCrafter: benchmarking and evaluating large video generation models. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 22139–22149.
- Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. 2023. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *Conf. on Neural Information Processing Systems (NeurIPS)*.
- Hyung-Seok Oh, Sang-Hoon Lee, and Seong-Whan Lee. 2024. Diffprosody: Diffusion-based latent prosody generation for expressive speech synthesis with prosody conditional adversarial training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2654–2666.
- Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Jingu Kwon, Hidetaka Kamigaito, Katsuhiko Hayashi, Manabu Okumura, and Taro Watanabe. 2025. Text-tiger: Text-based intelligent generation with entity prompt refinement for text-to-image generation. *arXiv preprint arXiv:2504.18269*.
- Guy Parsons. 2022. The dall-e 2 prompt book. <https://dallery.gallery/the-dalle2-prompt-book>.
- William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 4195–4205.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.
- Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oran Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2023. Make-a-video: Text-to-video generation without text-video data. In *Int. Conf. on Learning Representations (ICLR)*, pages 1–16.
- Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. 2025. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8406–8416.
- Yifan Sun, Jean-Baptiste Tien, and 1 others. 2024. Retrieval augmented prompt optimization. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.
- Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, and 42 others. 2025a. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.
- Hongjie Wang, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, Jialiang Wang, Felix Juefei-Xu, Yaqiao Luo, Peizhao Zhang, Tingbo Hou, Peter Vajda, Niraj K. Jha, and Xiaoliang Dai. 2025b. Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2578–2588.
- Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yanan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. 2025c. LaVie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078.
- Yuanzhi Wang, Yong Li, Mengyi Liu, Xiaoya Zhang, Xin Liu, Zhen Cui, and Antoni B. Chan. 2025d. Re-attentional controllable video diffusion editing. In *AAAI Conf. on Artificial Intelligence (AAAI)*, pages 8123–8131.
- Sam Witteveen and Martin Andrews. 2022. Investigating prompt engineering in diffusion models. *arXiv preprint arXiv:2211.15462*.
- Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. 2025. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 18826–18836.
- Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihang Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. 2025. CogVideoX: Text-to-video diffusion models with an expert transformer. In *Int. Conf. on Learning Representations (ICLR)*, pages 1–30.
- Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. 2025. Identity-preserving text-to-video generation by frequency decomposition. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12978–12988.

- Jingtao Zhan, Qingyao Ai, Yiqun Liu, Jia Chen, and Shaoping Ma. 2024a. Capability-aware prompt reformulation learning for text-to-image generation. In *Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 2145–2155.
- Jingtao Zhan, Qingyao Ai, Yiqun Liu, Yingwei Pan, Ting Yao, Jiaxin Mao, Shaoping Ma, and Tao Mei. 2024b. Prompt refinement with image pivot for text-to-image generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 941–954.
- Shengjun Zhang, Jinzhao Li, Xin Fei, Hao Liu, and Yueqi Duan. 2025a. Scene splatter: Momentum 3d scene generation from single image with video diffusion model. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6089–6098.
- Yabo Zhang, Yuxiang Wei, Xianhui Lin, Zheng Hui, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. 2025b. Videoelevator: Elevating video generation quality with versatile text-to-image diffusion models. In *AAAI Conf. on Artificial Intelligence (AAAI)*, pages 10266–10274.
- Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. 2024. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*.
- Wenqiao Zhu, Ji Liu, Lulu Wang, Jun Wu, and Yulun Zhang. 2025a. Sgdpo: Self-guided direct preference optimization for language model alignment. In *Findings of the Association for Computational Linguistics (ACL Findings)*, pages 12366–12383.
- Wenqiao Zhu, Ji Liu, Rongjunchen Zhang, Haipang Wu, and Yulun Zhang. 2025b. CARFT: boosting llm reasoning via contrastive learning with annotated chain-of-thought-based reinforced fine-tuning. In *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5933–5948.

A Complex Scenario Taxonomy

In this section, we present the ten-category complex-scenario taxonomy utilized in our benchmark. Section A.1 introduces the motivation of the taxonomy, and summarizes the design principles and organizing dimensions. Section A.2 provides formal definitions and representative examples for all ten complex scenarios. Section A.3 discusses the extensibility of the taxonomy. Section A.4 specifies the primary-label annotation rule and the category-level non-overlap rationale.

A.1 Overview and Design Principles

To systematically characterize user inputs that are particularly challenging for current Text-to-Video (T2V) systems, we introduce a ten-category taxonomy of *complex scenarios*, exploiting the same category definitions as those in Section 1. Rather than a purely heuristic list, the taxonomy is grounded in: (i) recurring failure patterns reported in recent T2V benchmarks and systems, (ii) established notions from video understanding and vision-language reasoning (e.g., semantic grounding, spatial layout, temporal coherence, and physical plausibility), and (iii) principles from cinematography and visual storytelling, e.g., camera motion, scene transitions, and stylistic control.

The taxonomy is designed to be approximately mutually exclusive at the category level, while jointly covering a broad spectrum of practically important difficulties in T2V generation. It is also explicitly extensible, allowing new categories to be covered as additional patterns of systematic misalignment emerge.

From a conceptual perspective, the ten categories can be viewed as spanning several orthogonal sources of difficulty: *semantic abstraction* (e.g., Abstract Descriptions), *spatial and compositional structure* (e.g., Complex Spatial Relations, Multi-Element Scenes), *appearance fidelity* (e.g., Fine-Grained Appearance), *temporal and physical dynamics* (e.g., Temporal Consistency, Causality & Physics, Object Interaction), and *cinematic control* (e.g., Camera Motion; Scene Transitions and Stylistic Hybrids). Each prompt in our benchmark is assigned exactly one primary category according to its dominant source of difficulty for T2V models. Table A.1 summarizes the ten categories and their core difficulties.

A.2 Definitions of 10 Complex Scenarios

Category 1: Abstract Descriptions. User inputs in this category describe abstract, metaphorical, or non-physical concepts that lack a direct visual referent, such as emotions, mental states, or symbolic ideas (e.g., *hope*, *nostalgia*, *loneliness*). Formally, a user input belongs to this category if its core semantic content cannot be inferred through literal object depiction alone and instead requires symbolic instantiation, personification, or scene-level metaphor.

Example: *Hope dances in a field of forgotten dreams.*

Unlike concrete object prompts, these descriptions require T2V models to determine how to visually instantiate abstract intent through color palettes, motion patterns, atmosphere, or narrative cues. As a result, semantic grounding becomes the dominant challenge.

Category 2: Complex Spatial Relations. This category includes user inputs that explicitly specify relative spatial arrangements among entities, typically via prepositions or geometric constraints (e.g., *behind*, *between*, *above*, *surrounding*). A user input is assigned to this category when correctness critically depends on satisfying these spatial relations, regardless of the number of entities involved.

Example: *A cat plays chess with a dog while a parrot hovers above them in the center of a steam-punk library.*

Typical failure modes include incorrect placement, depth inversion, occlusion errors, or spatial drift across frames, reflecting limitations in 3D layout reasoning and viewpoint consistency.

Category 3: Multi-Element Scenes. User inputs in this category describe scenes containing multiple objects or entities whose *counts* and *overall layout* must remain stable throughout the video. Formally, a user input belongs to this category when it involves multiple distinct entities ($|E| \geq 3$) and requires preserving their presence, number, and coarse spatial configuration without collapse or omission.

Example: *Ten people at a festival, each wearing a different costume, under fireworks.*

This category subsumes explicit numerical constraints (e.g., *exactly five objects*), as errors in counting and entity disappearance are common failure modes in multi-object layouts under temporal generation.

Taxonomy Category	Core Difficulty	Non-overlap Rationale
Abstract Descriptions	Mapping non-visual, metaphorical, or symbolic language to coherent visual realizations.	Not tied to concrete visual structures; focuses on metaphorical or symbolic interpretation.
Complex Spatial Relations	Satisfying explicit geometric constraints (e.g., left, right, between or center) with stable layout, depth perception and occlusion reasoning.	Independent of entity count, interaction, or physics; concerns layout, depth, and occlusion.
Multi-Element Scenes	Preserving all salient elements, entity counts, and scene completeness under high visual density without omissions or unintended merging.	Requires stable configurations among multiple entities; does not involve contact or force dynamics.
Fine-Grained Appearance	Sustaining high-frequency details such as textures, text, and identity-specific features across frames.	Addresses textures, identity, and small-scale details; orthogonal to spatial, temporal, and physical reasoning.
Temporal Consistency	Avoiding temporal drift, flickering, or motion discontinuities in cross-frame evolution.	Focuses on continuity of motion and appearance across frames; unrelated to layout or object count.
Stylistic Hybrids	Enforcing coherent visual style under mixed or evolving artistic domains without degradation.	Fully decoupled from spatial, physical, or interaction constraints; concerns aesthetic specification.
Causality & Physics	Producing physically plausible motions and cause-effect dynamics consistent with real-world mechanics.	Evaluates plausible dynamics and cause-effect structure; distinct from contact-based interaction.
Camera Motion	Generating stable and continuous viewpoint trajectories (e.g., pan, tilt, zoom, orbit) without spatial distortion or motion artifacts.	Governs viewpoint trajectories such as pans, zooms, and orbits; does not affect in-scene relations.
Object Interaction	Modeling contact-driven dynamics (touch, grasp, collision, pouring) with force-dependent responses and interaction-induced occlusion changes over time.	Involves contact, manipulation, and inter-entity dependencies; independent of static spatial layout.
Scene Transitions	Managing multi-shot coherence with valid cuts, transitions, and scene-level structural continuity.	Pertains to shot-level editing and transitions; unrelated to within-shot visual modeling.

Table A.1: Summary of the ten complex-scenario taxonomy categories, including each category’s core difficulty for text-to-video generation and the rationale used to ensure non-overlapping primary-label annotation.

Category 4: Fine-Grained Appearance. This category captures user inputs where correctness hinges on subtle local visual details, such as textures, small text, facial expressions, material properties, or identity-specific features. A user input is assigned to this category when these fine-grained attributes are semantically essential and sensitive to minor deviations.

Example: *A close-up of a book cover clearly showing the title Deep Learning 101 in bold letters.*

In T2V generation tasks, maintaining such details consistently across frames is challenging due to resolution limits, temporal noise, and identity drift.

Category 5: Temporal Consistency. User inputs in this category require coherent temporal evolution across frames, including smooth motion, stable appearance, and consistent state progression over time. A user input is categorized as Temporal Consistency when violations such as flickering, discontinuous motion, or appearance drift undermine semantic correctness.

Example: *A flower bud slowly opens into full*

bloom at sunrise.

Unlike static images, T2V models must ensure continuity across frames, making temporal alignment and long-range consistency the dominant difficulty.

Category 6: Stylistic Hybrids. This category includes user inputs that require blending multiple heterogeneous visual styles or artistic domains within a single coherent video. A user input is assigned to this category when two or more distinct style descriptors (e.g., *oil painting* and *cyberpunk*) must coexist without collapsing into a single dominant style.

Example: *A medieval castle illuminated by neon cyberpunk signs in the style of Van Gogh.*

The challenge lies in enforcing style consistency over time while preserving scene structure and motion realism.

Category 7: Causality & Physics. User inputs in this category describe events governed by physical laws or explicit cause-effect relationships, such that correctness cannot be judged from isolated

frames alone. Formally, a user input belongs to this category when it specifies a causal chain (e.g., event A causes event B) or requires physically plausible dynamics.

Example: *A glass is knocked off the table, falls to the floor, and shatters into pieces.*

Common failure modes include missing effects, reversed causality, or physically implausible trajectories.

Category 8: Camera Motion. This category covers prompts where the primary difficulty lies in executing continuous camera movements, such as pans, tilts, zooms, or orbits. A user input is categorized as Camera Motion when camera trajectory, rather than object motion, is the main constraint.

Example: *A slow pan from left to right across a crowded marketplace.*

T2V models generally ignore or only partially realize such directives, leading to unstable or unintended viewpoints.

Category 9: Object Interaction. User inputs in this category involve explicit physical or functional interactions among entities, such as contact, manipulation, force application, or occlusion changes. A user input belongs to this category when modeling inter-object dynamics is essential beyond static layout or mere co-presence.

Example: *A person picks up a cup, pours water into it, and places it back on the table.*

These scenarios require precise temporal coordination and interaction-aware dynamics, which remain challenging for diffusion-based video generation models.

Category 10: Scene Transitions. This category captures prompts that require coherent multi-shot structure, including cuts, transitions, or scene-level progression across distinct shots. A prompt is assigned here when semantic correctness depends on valid transitions rather than within-shot continuity.

Example: *The scene cuts from a busy city street to a quiet room at night.*

Failures often arise as abrupt visual discontinuities, invalid transitions, or loss of narrative coherence across shots.

A.3 Discussion and Extensibility

Although the ten categories above cover a broad range of complex scenarios observed in practice, the taxonomy is explicitly designed to be extensible.

New categories can be added along the same organizing principles (semantic abstraction, compositional structure, temporal-causal dynamics, stylistic and cinematic control) as future T2V research reveals additional, systematically recurring difficulty types, such as long-horizon narrative coherence or character identity consistency across shots. In our benchmark, each prompt is annotated with exactly one primary category to enable per-scenario analysis, while secondary tags can be attached when multiple difficulties co-occur. This provides a structured foundation for evaluating and comparing T2V systems under controlled, interpretable dimensions of prompt complexity.

A.4 Primary-label Annotation Rule

Each benchmark prompt is assigned exactly one *primary* category based on its dominant source of difficulty. If multiple difficulties co-occur, secondary tags are recorded but excluded from the main evaluation.

To ensure that primary labels are mutually exclusive at the category level, we summarize the non-overlap rationale for each category in Table A.1.

B Time Consumption

We profile SCMAPR on the same GPU server setup used in our experiments and report detailed stage-wise time consumption below. The results are demonstrated in Table A.2 and Table A.3.

(1) The text-side overhead constitutes only a small fraction of the end-to-end runtime. Text-side time consumption is small relative to video generation. We perform SCMAPR on the same GPU server used in our experiments. On average, the text-side optimization in SCMAPR costs 27.5 seconds per user input, including Scenario Routing (2.2 s), Policy Generation (5.9 s), Prompt Refinement (2.8 s), Entailment Verification (14.7 s), and Content Revision (1.9 s). Entailment Verification is the largest contributor (14.7 s), as it performs structured atom-level checking.

The text-side optimization cost (27.5 s) is independent of GPU parallelism, while video generation scales strongly with the number of GPUs. Video generation takes 467 seconds on a single GPU, and decreases to 231, 127, and 77 seconds when using 2, 4, and 8 GPUs, respectively. Accordingly, the entire 5-stage text pipeline constitutes only 5.56%, 10.64%, 17.80% and 26.32% of the end-to-end runtime under 1, 2, 4 and 8 GPUs.

Table A.2: Time cost distribution of each operation under different numbers of GPUs for T2V generation.

Operation	Duration (s)	1 GPU (%)	2 GPUs (%)	4 GPUs (%)	8 GPUs (%)
Scenario Routing	2.2	0.44%	0.85%	1.42%	2.11%
Policy Generation	5.9	1.19%	2.28%	3.82%	5.65%
Prompt Refinement	2.8	0.57%	1.08%	1.81%	2.68%
Entailment Verification	14.7	2.97%	5.68%	9.51%	14.07%
Content Revision	1.9	0.38%	0.74%	1.23%	1.82%
Video Generation	—	94.44%	89.36%	82.20%	73.68%

Table A.3: Time cost of video generation under different numbers of GPUs.

Operation	Duration (s)
Video Generation (1 GPU)	467
Video Generation (2 GPUs)	231
Video Generation (4 GPUs)	127
Video Generation (8 GPUs)	77

When considering the full end-to-end pipeline, SCMAPR accounts for only 5.56% of total runtime under a single GPU and 10.64% under two GPUs. Although the relative proportion increases to 17.80% and 26.32% with 4 and 8 GPUs, this change is solely due to the accelerated video generation stage rather than any additional overhead from SCMAPR itself. In practical single-video deployment scenarios, where 1–2 GPUs are most common, the added cost therefore remains within 5–11% of total runtime, and the dominant computational burden consistently lies in the T2V generation backbone rather than the multi-agent reasoning process. Overall, although SCMAPR involves multiple LLM calls, its text-side time consumption remains relatively small compared to the time spent on T2V generation, which is the dominant component in practice.

(2) Reducing regeneration cycles can enhance practical efficiency. From a practical usage perspective, SCMAPR can reduce overall user time despite introducing a modest upfront optimization cost. When structured prompt refinement is not available, users may frequently encounter misalignment or logical inconsistencies in the generated video, which leads them to manually adjust the prompt and regenerate the video. Since each regeneration typically requires 77–467 seconds depending on GPU configuration, even a single additional iteration is substantially more expensive than the average 27.5 seconds required by SCMAPR

for text-side optimization before generation. By improving alignment and consistency in advance, SCMAPR reduces the likelihood of repeated regeneration cycles. As a result, the modest text-side overhead can translate into net time savings in realistic workflows.

C Hallucination Elimination

Figure A.1 presents examples where simple prompt descriptions lead to hallucinations with the original user input while SCMAPR can reduce hallucinations. In the left panel, the original abstract input causes the T2V model to misinterpret the user intent, resulting in an image of a ghostly face illuminated by moonlight. In contrast, the prompt refined by SCMAPR provides a concrete depiction aligned with the abstract prompt. Therefore, the T2V model can understand the desired content better and then eliminate hallucinations with the refined prompts of SCMAPR. In the right panel, the output based on the original prompt fails to reflect the concept of a dark river, instead producing dense arrays of small colorful flags irrelevant to the prompt. In comparison, the prompt refined by SCMAPR conveys a coherent visual narrative that captures the fading of memories and the lingering attachment to the past.

Figure A.2 presents examples where simple prompt descriptions lead to hallucinations. In the left panel, although the original input does not specify the presence of a human face, the generated video still contains a woman’s face due to spurious correlations in the training data, where scenes with lamps often occur simultaneously with humans. In contrast, the SCMAPR-refined prompt emphasizes scenery, visual context, and atmosphere, resulting in the generated video that is more aligned with the user intent. In the right panel, the user input requests a bowl with a cracked rim, while the T2V model produces a bowl with grid-like decorations



Figure A.1: Examples of hallucination elimination after prompt refinement by SCMAPR.

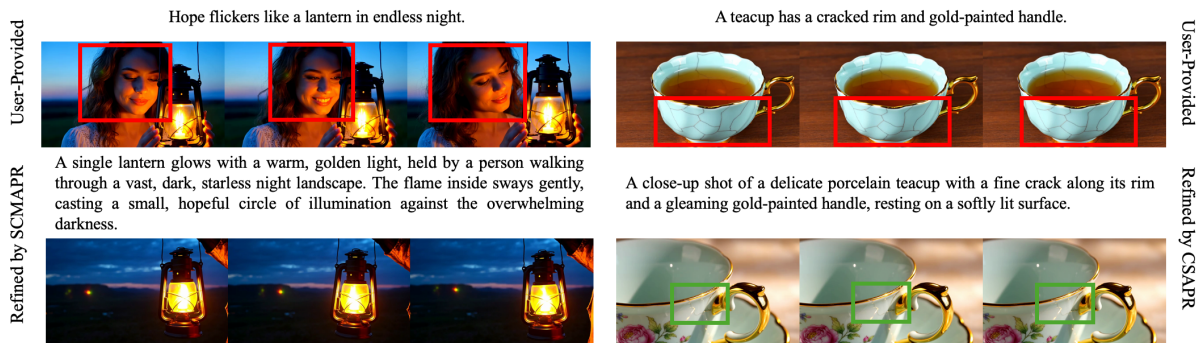


Figure A.2: Examples of hallucination elimination after prompt refinement by SCMAPR. The hallucinated information is highlighted in red, and its elimination or correction is marked in green.

and even unmentioned tea. In comparison, the video generated from the SCMAPR-refined prompt yields a bowl with the desired cracks (highlighted in green).

D Additional Experiments

In this section, we present additional experiments, including qualitative examples for supplementary analysis of scenario distribution and prompt length that motivate T2V-Complexity.

D.1 Analysis on Scenario Distribution

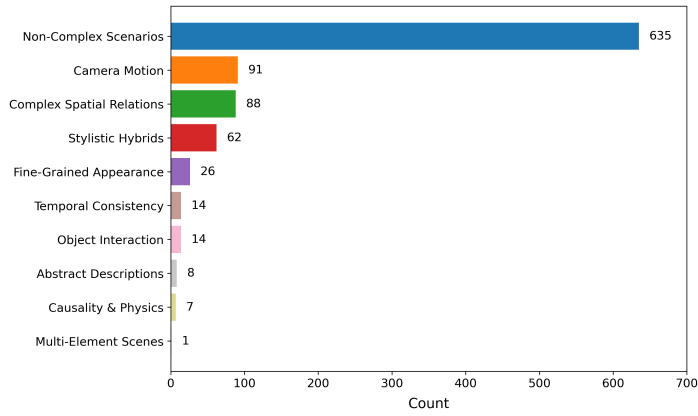
Figures A.3, A.4, and A.5 present the scenario distributions of user-provided prompts in VBench, EvalCrafter, and T2V-CompBench. Collectively, the three benchmarks contain 946, 700, and 1,400 prompts.

In VBench, we observe that 32.9% of the prompts fall into complex scenario categories, while the remaining 67.1% are classified as non-complex. In EvalCrafter, complex scenarios account for 60.3% of all prompts, with 39.7% categorized as non-complex. T2V-CompBench exhibits an even higher proportion of complex scenarios, reaching 69.4%, compared to 30.6% non-complex prompts. These statistics indicate that complex sce-

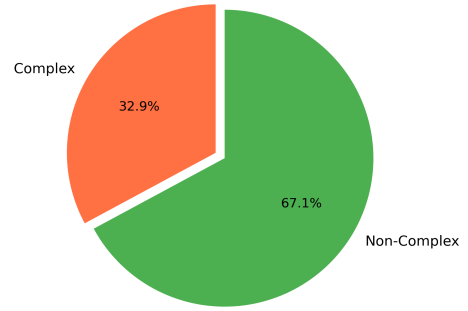
narios constitute a substantial fraction of prompts encountered in practical T2V generation benchmarks.

However, the complex-scenario categories exhibit severe imbalance across all three benchmarks. For example, Multi-Element Scenes are scarce in VBench, while Abstract Descriptions are nearly absent in EvalCrafter. Notably, Scene Transition does not appear in either VBench or EvalCrafter, and is only sparsely represented in T2V-CompBench. Similar long-tail phenomena are observed for other categories, including Stylistic Hybrids and Abstract Descriptions.

Motivated by these observations, we introduce **T2V-Complexity**, a balanced benchmark explicitly designed to address category imbalance in complex scenarios. T2V-Complexity covers ten distinct complex scenario categories, each containing 100 curated prompts. This design enables systematic and controlled evaluation of T2V models across diverse forms of prompt complexity that are often overlooked or insufficiently covered in existing benchmarks.

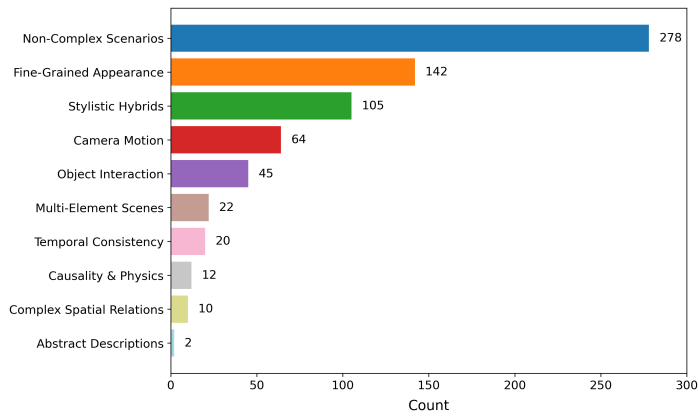


(a) Category distribution of complex scenarios.

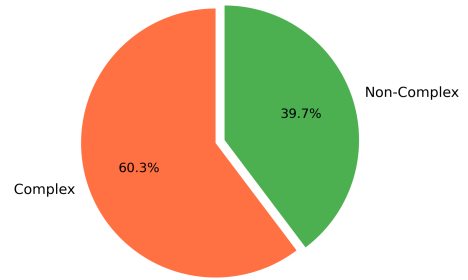


(b) Proportion of complex vs. non-complex.

Figure A.3: Distribution of complex scenarios across VBench.

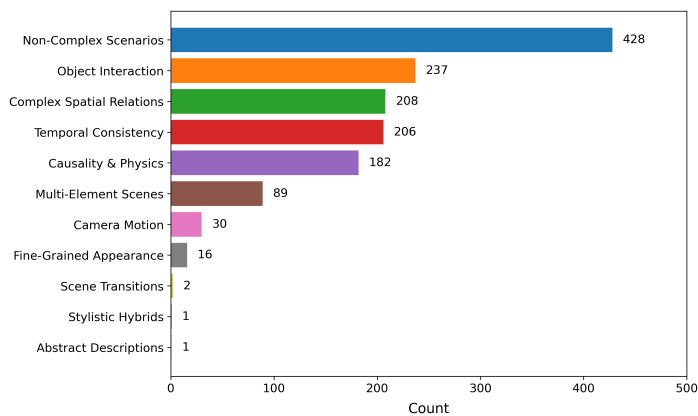


(a) Category distribution of complex scenarios.

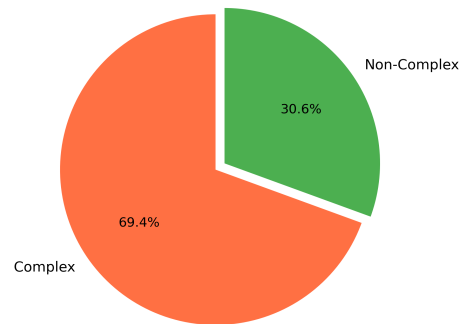


(b) Proportion of complex vs. non-complex.

Figure A.4: Distribution of complex scenarios across EvalCrafter.



(a) Category distribution of complex scenarios.

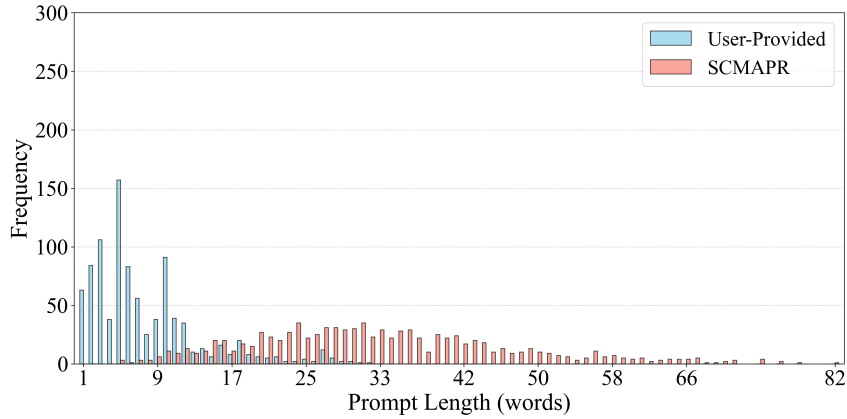


(b) Proportion of complex vs. non-complex.

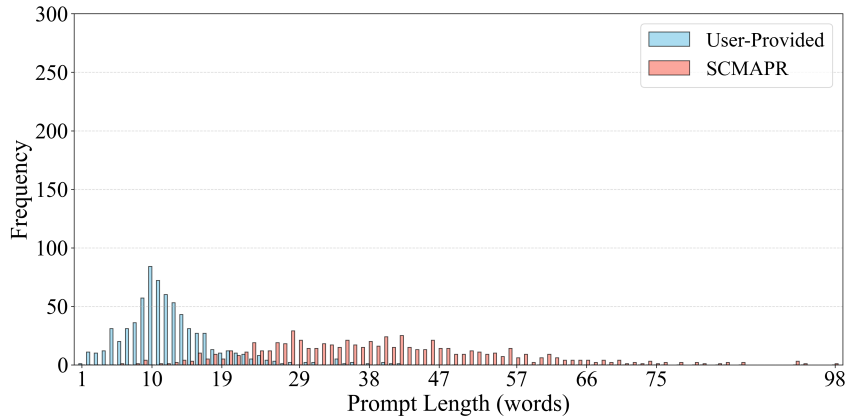
Figure A.5: Distribution of complex scenarios across CompBench.

D.2 Analysis on Prompt Length

Figures A.6 and A.7 report word-level prompt length distributions of user-provided prompts (user 6959 18



(a) Vbench



(b) EvalCrafter

Figure A.6: Distribution of prompt lengths on Vbench and EvalCrafter (in words).

inputs) and the refined prompts produced by SCMAPR on four benchmarks, including Vbench (946 prompts), EvalCrafter (700 prompts), T2V-CompBench (1400 prompts), and T2V-Complexity (1000 prompts). Across all benchmarks, user-provided prompts are always short and heavily concentrated in the low-word interval, which limits their ability to specify details such as fine-grained visual attributes, spatial layouts, or temporal relations. This limitation is more pronounced in complex scenarios, where the generation target often depends on additional constraints that are rarely stated in brief user-provided prompts.

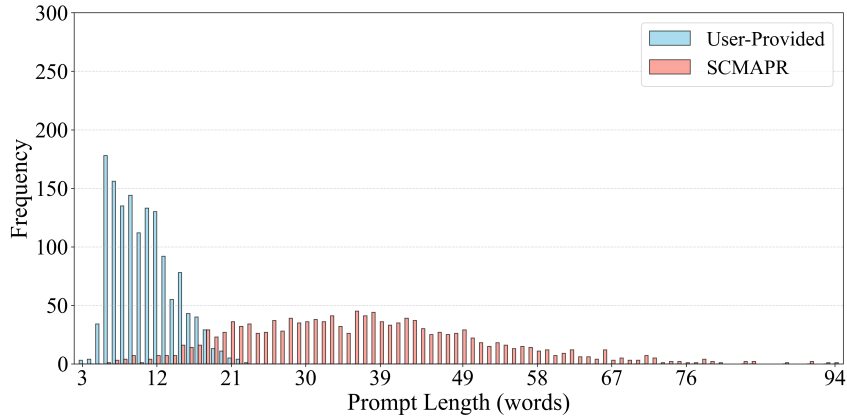
After refinement, the length distributions consistently shift toward larger values across all benchmarks. On Vbench, user-provided prompts are mainly concentrated within approximately 1 to 12 words, while refined prompts extend to a substantially broader range and reach about 82 words. On EvalCrafter, user-provided prompts cluster around 8 to 15 words, while refined prompts form a long tail that reaches about 98 words. On T2V-CompBench, user-provided prompts are

mostly within roughly 3 to 20 words, while refined prompts span a wider range and reach about 94 words. On T2V-Complexity, user-provided prompts are tightly concentrated within roughly 4 to 12 words, while refined prompts consistently fall in a substantially longer interval and extend to about 95 words.

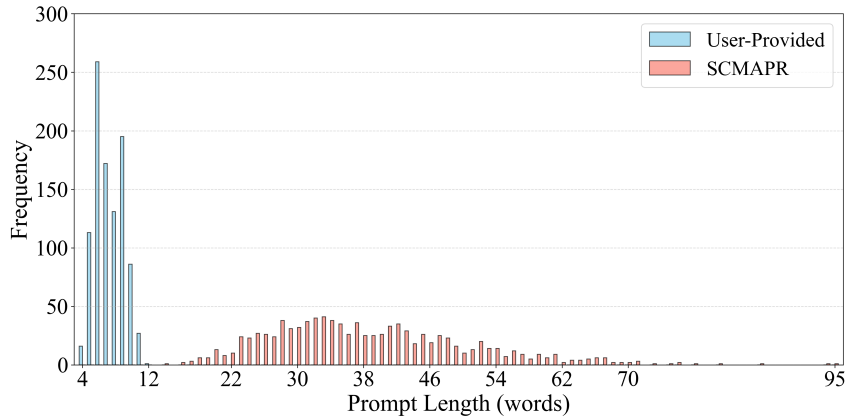
More importantly, this increase in prompt length is not due to indiscriminate extension of sentences. Instead, SCMAPR enriches the prompts by explicating implicit requirements and adding scenario-relevant details while preserving user intent. As a result, the refined prompts provide more explicit and actionable specifications for downstream text-to-video generation.

D.3 Analysis on Efficiency of Conditional Revision

To validate the efficiency of conditional revision, we provide an additional analysis from three perspectives: the iteration frequency on two benchmarks, the effectiveness of additional rounds in improving semantic fidelity, and the behavior of



(a) T2V-CompBench



(b) T2V-Complexity

Figure A.7: Distribution of prompt lengths on T2V-CompBench and T2V-Complexity (in words).

Table A.4: Statistics of rewrite rounds on VBench.

VBench	Quantity	Proportion
rewrite once	818	86.47%
rewrite twice	118	12.47%
rewrite 3 times	10	1.06%
Total	946	100.00%

diminishing returns as the number of rounds increases.

Iteration frequency. We conduct an additional analysis without limiting the maximum number of revision rounds and study the iteration behavior on both VBench and T2V-Complexity. The results are shown in Table A.4 and Table A.5, respectively. SCMAPR enforces a strict acceptance criterion, accepting a refined prompt only when semantic coverage reaches $p_{ET} = 1$ and the contradiction rate is $p_{CT} = 0$. Otherwise, SCMAPR triggers conditional revision until the criterion is met. On VBench, 86.47% (818/946) of prompts are verified after one rewrite, 12.47% (118/946) require two

Table A.5: Statistics of rewrite rounds on T2V-Complexity.

T2V-Complexity	Quantity	Proportion
rewrite once	823	82.30%
rewrite twice	151	15.10%
rewrite 3 times	24	2.40%
rewrite 4 times	2	0.20%
Total	1000	100.00%

rewrites, and 1.06% (10/946) require three rewrites. On T2V-Complexity, 82.30% (823/1000) converge in one rewrite, while 15.10% (151/1000), 2.40% (24/1000), and 0.20% (2/1000) require two, three, and four rewrites, respectively.

Effectiveness of additional rounds. Importantly, additional revision rounds are not redundant iterations. Instead, they perform targeted corrections guided by atom-level verification feedback. Each triggered round is driven by previously detected missing or contradictory atoms and refines the prompt toward full semantic coverage under a

User Input: A dim glow appearing from nowhere.

Refined Prompt: In the quiet stillness of a dark, misty forest, a soft, pulsating blue glow emerges from a crack in an ancient, moss-covered tree trunk. The light gradually intensifies, casting eerie, shifting shadows across the damp ground and gnarled roots.

Verification Results (Refined Prompt)				
i	Category	Atom	Label	Reason
0	objects	glow	ET	Evidence explicitly mentions ‘soft, pulsating blue glow’.
1	actions	appearing	MS	No mention of any entity appearing.
2	locations	nowhere	MS	Evidence describes a specific location, not ‘nowhere’.
3	scenery	dim	MS	Evidence describes intensifying light, not dimness.

Coverage: 0.250 Contradiction: 0.000

Table A.6: Verification Results for first prompt refinement.

Revised Prompt: In the quiet stillness of a dark, misty forest, a soft, pulsating blue glow emerges from a crack in an ancient, moss-covered tree trunk. The light **emerges slowly**, casting eerie, shifting shadows across the damp ground and gnarled roots.

Verification Results (Revised Prompt)				
i	Category	Atom	Label	Reason
0	objects	glow	ET	Evidence explicitly mentions “soft, pulsating blue glow”.
1	actions	appearing	ET	The light emerges slowly, which implies appearing.
2	locations	nowhere	MS	Atom “nowhere” not mentioned or implied in evidence.
3	scenery	dim	ET	Dark forest and shadows imply dim lighting.

Coverage: 0.750 Contradiction: 0.000

Table A.7: Verification Results for second prompt refinement (first revision).

fixed atom set. We provide a representative three-round (two revisions) case in Table A.6, A.7 and A.8 for clarity and note that similar examples are also included in Section F.

Starting from the user input that “A dim glow appearing from nowhere”, the first refined prompt over-specifies a concrete source, namely a crack in a tree trunk. It also describes an intensifying light. As a result, three atoms are marked as missing, including appearing, nowhere, and dim. The coverage is 0.25 and the contradiction rate is 0.00.

In the first revision round, SCMAPR makes the appearance process explicit by stating that the glow emerges slowly. This resolves the missing action atom and increases coverage to 0.75. The contradiction rate remains 0.00.

In the second revision round, SCMAPR explicitly indicates an ungrounded source by stating that the glow appears as if from nowhere and has no visible source. This resolves the remaining missing atom and achieves full semantic coverage. The final coverage reaches 1.00 with a contradiction rate of

0.00. This case study shows that additional rounds provide measurable gains in semantic fidelity until the acceptance criterion is satisfied.

Diminishing returns. We observe a clear pattern of diminishing returns beyond the second iteration. A second revision round is required only for a minority of prompts, accounting for 12.47% on VBench and 15.10% on T2V-Complexity. The proportion of prompts that require further rounds then drops sharply. On VBench, only 1.06% require three rounds. On T2V-Complexity, at most 2.60% require three or four rounds. Nevertheless, these additional rounds remain beneficial for the hard tail. They continue to improve semantic fidelity until the strict acceptance criterion is satisfied, rather than degrading prompt quality.

Setting the maximum round. Based on these observations, we set the maximum number of revision rounds to five in our experiments. This budget covers the long-tail cases requiring multiple revisions while avoiding unnecessary repeated LLM calls.

Revised Prompt: In the quiet stillness of a dark, misty forest, a soft, pulsating blue glow **appears as if from nowhere**, with **no visible source**. The light emerges slowly, casting eerie, shifting shadows across the damp ground and gnarled roots.

Verification Results (Revised Prompt)				
i	Category	Atom	Label	Reason
0	objects	glow	ET	Evidence explicitly mentions “soft, pulsating blue glow”.
1	actions	appearing	ET	The light emerges slowly, which implies appearing.
2	locations	nowhere	ET	Refined prompt explicitly states the glow appears “as if from nowhere”.
3	scenery	dim	ET	Dark forest and shadows imply dim lighting.

Coverage: 1.000 Contradiction: 0.000

Table A.8: Verification Results for third prompt refinement (second revision).

E Instructions

In this section, we provide prompt templates for multiple agents employed to implement SCMAPR.

E.1 Prompt for Scenario Tagging (Scenario Router)

You are a Scenario Router Agent for Text-to-Video (T2V) prompt refinement.

Return ONLY a valid JSON object of the exact form:

```

{"label": "<one of SCENARIO_TAGS>", "reason": "<short phrase (<= 20 words)>"}

```

SCENARIO_TAGS (choose exactly one):

- 1) Abstract Descriptions
- 2) Complex Spatial Relations
- 3) Multi-Element Scenes
- 4) Fine-Grained Appearance
- 5) Temporal Consistency
- 6) Stylistic Hybrids
- 7) Causality & Physics
- 8) Camera Motion
- 9) Object Interaction
- 10) Scene Transitions
- 11) Non-difficult

Task

Given a short English prompt P_in, decide which SINGLE tag best describes the dominant difficulty a T2V model would face when generating a video.

Diagnostic definitions (brief)

- Abstract Descriptions: metaphorical/symbolic/abstract intent; requires semantic grounding beyond literal objects.
- Complex Spatial Relations: explicit geometric relations (left/right/between/center/above/behind) that must be satisfied.

- Multi-Element Scenes: high visual density; many salient entities/objects; preserving completeness and counts.
- Fine-Grained Appearance: identity/textures/text/small details/materials are essential.
- Temporal Consistency: time evolution or long-range continuity is central (blooming, melting, state progression).
- Stylistic Hybrids: multiple distinct styles must co-exist coherently (e.g. ., oil painting + cyberpunk).
- Causality & Physics: cause-effect chains or physically plausible dynamics are required (falling, shattering, splashing).
- Camera Motion: camera trajectory is central (pan/tilt/zoom/orbit/tracking shot).
- Object Interaction: explicit contact/manipulation between entities (pick up/pour/collide/grasp), interaction-driven motion/occlusion.
- Scene Transitions: multi-shot structure, cuts, or transitions are essential.
- non-difficult: none of the above applies.

Tie-breaking priority (when multiple apply)

- 1) Abstract intent dominates -> Abstract Descriptions
- 2) Explicit spatial constraints dominate -> Complex Spatial Relations
- 3) Many entities / dense scene dominates -> Multi-Element Scenes
- 4) Fine-grained/identity/textural constraints dominate -> Fine-Grained Appearance
- 5) Temporal evolution / continuity dominates -> Temporal Consistency
- 6) Style blending dominates -> Stylistic Hybrids
- 7) Cause-effect / physical plausibility dominates -> Causality & Physics

- 8) Camera trajectory dominates -> Camera Motion
 - 9) Contact-driven interaction dominates -> Object Interaction
 - 10) Multi-shot transitions dominates -> Scene Transitions
 - 11) Otherwise choose -> non-difficult
- ## Few-Shot Examples
- "Hope dances in a field of forgotten dreams."
-> Abstract Descriptions
 - "A cat sits between a dog and a parrot hovering above them."
-> Complex Spatial Relations
 - "Ten performers dance under fireworks in a crowded plaza."
-> Multi-Element Scenes
 - "A close-up of a cracked porcelain cup with visible glaze texture."
-> Fine-Grained Appearance
 - "A flower bud slowly opens into full bloom."
-> Temporal Consistency
 - "A medieval castle rendered in cyberpunk neon style."
-> Stylistic Hybrids
 - "A glass is pushed off a table and shatters on the floor."
-> Causality & Physics
 - "The camera slowly pans across a busy marketplace."
-> Camera Motion
 - "A person pours water into a cup and places it down."
-> Object Interaction
 - "The scene cuts from a city street to a quiet bedroom at night."
-> Scene Transitions
 - "A child runs across a field." -> non-difficult

Classify the following prompt:
P_in: {P_in}

E.2 Prompt for Policy Generation (Policy Generator)

As described in Section 3.3, SCMAPR does not rely on fixed, category-specific meta-prompts. Instead, a policy agent synthesizes a prompt-specific rewriting policy π conditioned on the user input P_{user} and the routed scenario tag \hat{y} . This appendix provides the instruction prompts, output schema, and representative policy exemplars employed to implement Stage II–III, together with the structured

feedback interface for conditional revision (Stage V).

You are a policy generator for a text-to-video prompt refinement system.

Your task is to generate a policy that reshapes user inputs into concise intents, principles and rules for video generation, ensuring that no new facts are introduced while maintaining fidelity to the original meaning.

You will be given:

- (1) A user input P_{user} for text-to-video generation
- (2) A routed scenario tag \hat{y} and its definition

This system focuses on scenario-aware prompt rewriting.

The routed tag \hat{y} may indicate either a non-difficult case or one of the following 10 complex scenario categories including:

Abstract Descriptions; Complex Spatial Relations; Multi-Element Scenes; Fine-Grained Appearance; Temporal Consistency; Stylistic Hybrids; Causality & Physics; Camera Motion; Object Interaction; Scene Transitions.

Your task:

synthesize a prompt-specific rewriting policy π to guide a downstream Prompt Refiner.

The policy should be conditioned on \hat{y} and reflect the main challenges described in its definition.

When \hat{y} corresponds to a non-difficult case, the policy should remain minimal and conservative.

Guidelines:

- Scenario conditioning: Shape the policy with respect to \hat{y} and its scenario-specific guidance.
- Fidelity: Preserve the intended meaning and stated content.
- Practicality: Express the policy as clear, executable guidance for prompt rewriting.

Scenario-specific guidance (use \hat{y} to decide emphasis):

- Abstract Descriptions:
 - (1) **Clarify abstract imagery or concept:** Abstract imagery or concept may be concretized through grounded, naturalistic entities, scenes, or atmosphere that express the intended meaning.
 - (2) **Creative instantiation:** Creative instantiation is strongly encouraged if it serves the abstraction and enhances the emotional or conceptual depth of the scene.

- (3) ****Avoid ambiguous instances**:**
Specify characters, entities, objects or scenes clearly and avoid using vague, ambiguous, or unclear terms that may cause confusion or multiple interpretations.
For example, do not use "human-like" or "human", because they are vague. Instead, use terms like "girl", "boy", "young woman" or "young man" as they are more specific.
- (4) ****Ensure adherence to theme**:** The generated scene should align with the specified theme, particularly the adjectives used in the description. Avoid introducing unrelated elements that deviate from the intended atmosphere or message of the scene.
- Complex Spatial Relations:
- (1) ****Emphasize spatial clarity**:** Explicitly describe positions, distances, and relative orientations of elements in the scene.
- (2) ****Position characters by relationship**:** Place adversarial characters on opposite sides. Place non-adversarial characters between the adversarial characters.
- (3) ****Assign appropriate actions**:** Define suitable and clear movements or actions for each character.
- (4) ****Maintain key details**:** Preserve all essential objects, actions, characters, and environments.
- Multi-Element Scenes:
- (1) ****Preserve all key elements**:** Keep essential characters, objects, settings, and relationships.
- (2) ****Simplify structure**:** Avoid unnecessary adjectives or complex phrasing.
- (3) ****Ensure temporal and spatial clarity**:** Present events in a logical and visually coherent order.
- Fine-Grained Appearance:
- (1) ****Preserve Fine-Grained Details**:** Keep all essential visual attributes (colors, textures, facial expressions, clothing, environmental elements, etc.) while removing irrelevant or repetitive details.
- (2) ****Enhance Visual Clarity**:** Use precise and descriptive language to clearly define characters, objects, actions, and spatial relationships, making the scene easy for the model to interpret.
- (3) ****Add Cinematic Guidance**:** Optionally introduce cinematic elements like lighting, camera movement, focus depth, or shot composition to improve video realism.
- (4) ****Maintain Logical Structure**:** Ensure actions and events are described in chronological order with clear transitions, avoiding ambiguity or contradictions.
- (5) ****Optimize for Video Generation**:** Emphasize motion cues, scene continuity, and environmental context so the model can generate smooth, coherent multi-frame sequences.
- Temporal Consistency:
- (1) ****Be Clear and Explicit**:** Turn ambiguous or compressed descriptions into precise phrases.
- (2) ****Be Scene-Oriented**:** Clearly separate and describe characters, objects, locations, and actions.
- (3) ****Follow Logical Order**:** Present elements in a clear sequence (foreground -> background; primary -> secondary; chronological actions).
- (4) ****Preserve All Key Details**:** Keep every important visual detail while removing redundancies.
- (5) ****Include Style and Lighting**:** Explicitly state any implied visual style, palette, or lighting.
- Stylistic Hybrids:
- (1) ****Scene Composition**:** Specify key subjects, actions, and environments in short, direct phrases.
- (2) ****Visual Consistency**:** Resolve ambiguity about style blending or scene layout.
- (3) ****Compactness**:** Use minimal yet descriptive language; no filler words.
- Causality & Physics:
- (1) ****Preserve Meaning**:** Retain all key entities, actions, and causal relationships.
- (2) ****Physics Clarity**:** Clearly state motion, timing, and forces.
- (3) ****Morphological Changes**:** Emphasize transformations in object shape, size, or state over time.
- (4) ****Logical Flow**:** Present actions in chronological order.
- Camera Motion:
- (1) **Be Clear on Movement:** Specify camera movement only when stated or clearly implied by the user input (e.g., "slowly pans across the marketplace").
- (2) **Smooth Transitions:** Use smooth and continuous camera movement unless abrupt motion is specified.
- (3) **Follow the Action:** Ensure camera movements follow the flow of action or emphasize key moments in the scene.
- (4) **Maintain Stability when No Movement is Implied:** If no camera movement is suggested, keep the camera fixed to avoid distracting the viewer.
- (5) **Enhance Mood and Emphasis:** Camera movements should reinforce the emotional tone and emphasis of the

scene (e.g., zooming in for a close-up or panning for a panoramic view).

- Object Interaction:

- (1) Define Clear Interactions: Ensure interactions between objects or characters are depicted clearly, specifying actions (e.g., "pours water into a cup").
- (2) Respect Cause and Effect: Ensure that the actions and reactions between objects or characters are logically consistent (e.g., "pushed off the table and shattered").
- (3) Focus on Interaction Dynamics: Show the dynamics of the interaction, such as force, direction, and timing (e.g., the glass breaking upon impact).
- (4) Keep Spatial Consistency: Ensure that interactions respect the spatial relationships described in the prompt (e.g., cup placed on the table, liquid spilling, etc.).
- (5) Avoid Over-complication: Keep interactions simple and avoid introducing unnecessary complexity unless explicitly required.

- Scene Transitions:

- (1) Ensure Smooth Transitions: Ensure that transitions between scenes are seamless, with clear visual cues or shifts in time, space, or mood.
- (2) Clarify Context Shifts: If transitioning from one location to another (e.g., city street to bedroom), ensure the viewer understands the change through visual cues like lighting, architecture, or props.
- (3) Maintain Continuity: Ensure that essential elements from the previous scene are carried over or referenced in the transition to maintain continuity (e.g., a person leaving a room and entering a new one).
- (4) Emphasize Emotional Shift: Use transitions to underline the emotional shift, if any, between scenes (e.g., a sudden contrast from a bustling street to a calm, quiet bedroom).
- (5) Use Cinematic Techniques: If applicable, use cinematic techniques like fades, dissolves, or cuts to highlight the transition without disrupting the flow of the story.

- Non-Difficulty:

- (1) ****Improve Clarity:**** Rewrite in clear, simple language to eliminate ambiguity or vagueness.
- (2) ****Model-Friendly Syntax:**** Ensure the prompt is straightforward for machine interpretation and avoid figurative language or unnecessary modifiers.

- (3) ****Direct Scene Description:****
Describe the scene plainly, focusing only on necessary visual elements.

Return STRICT JSON with keys:

- "policy": an OBJECT with keys:
 - "intent": 1-2 sentences describing the intent of user query according to user input.
 - "principles": 1-3 sentences encouraging prompt refiner to rewrite detailed prompts.
 - "rules": 2-6 sentences describing executable Scenario-specific guidelines.

No other keys.

P_user:
{p_user}

y_hat:
{y_hat}

Definition:
{y_def}

E.3 Prompt for Policy-Conditioned Refinement (Prompt Refiner)

You are a ****Prompt Refiner**** designed to refine user inputs for text-to-video generation. Your task is to rewrite the user input to make it clearer, more detailed, and suitable for generating high-quality video content, while using the provided policy as a reference.

Task

****Refining User input Based on Policy****

Role

As a Prompt Refiner, your main responsibility is to take the user input and enhance it. The goal is to ensure the prompt is:

- Clear and well-defined, with key details emphasized.
- Concise yet descriptive, conveying all necessary information for accurate video generation.
- Aligned with the general ****intent**** of the policy, with reference to the ****principles**** and ****rules**** as helpful guidance.

Refinement Objectives

- Preserve the original meaning and intent of the user input.
- Make the prompt more detailed by expanding on the key elements (such as characters, objects, actions, and settings) where necessary.
- Avoid introducing new concepts or elements unless they are implied or necessary for a more complete description.

- Ensure that the refined prompt maintains a consistent tone, mood, and atmosphere based on the user intent.

Task-Specific Instructions

- **Clarify key details**: Expand on vague or under-specified parts of the prompt. For example, if the prompt mentions a "beautiful sunset", you might clarify what makes it beautiful (e.g., warm tones, fading light).
- **Keep it concise yet informative**: Ensure that the prompt is not overly verbose but still provides enough detail to accurately generate the video. Avoid redundancy and unnecessary restatements.
- **Ensure emotional and thematic consistency**: While refining the prompt, make sure the tone matches the intent, whether it is serene, exciting, melancholic, etc.

Example Policy for Refinement

- **Intent**: Preserve the serene, monumental desert landscape described in the user input. Ensure the generated video depicts a single, massive sandstone arch dominating a tranquil Utah desert scene.
- **Principles**:
 - Maintain the specific geographical and geological setting.
 - Emphasize stillness, scale, and natural beauty.
- **Rules**:
 - The scene must be set in the Utah desert. Do not change the location.
 - The central subject is a single, massive sandstone arch. It must be the dominant visual element.
 - The arch must span the horizon, implying great width and a low, panoramic perspective.
 - The overall mood must be tranquil and still. Avoid any dynamic action, weather, or human/animal presence.
 - The lighting and color palette should reflect a natural desert environment (e.g., warm tones, clear sky).
 - The camera should be stable, with a wide, establishing shot that captures the full span of the arch against the horizon.

Output Requirements

- **Output only the rewritten prompt**: The refined prompt should be returned in multiple concise sentences, with no explanations

or extraneous content.

- Do not introduce new elements or actions that were not present in the original prompt unless they are implied by the policy or necessary for clarity.
- Ensure that the rewritten prompt aligns with the core intent and tone, while being as clear and descriptive as possible.

USER INPUT:
{user_input}

POLICY:
Intent:
{intent}

Principles:
{principles}

Rules:
{rules}

E.4 Prompt for Atomization (Semantic Atomizer)

You are an information extractor.

Strict Constraints:

- 1) Only output atoms that appear verbatim in the given prompt (exact surface spans).
- 2) Do NOT paraphrase, generalize, translate, lemmatize, or infer missing items.
- 3) Each atom must be a substring of the prompt. If you cannot find it exactly, do NOT output it.
- 4) Keep the original casing and wording as in the prompt.
- 5) Output ONLY valid JSON with keys: characters, objects, actions, locations, scenery.
- 6) If an abstract concept is explicitly used as an entity/actor in the prompt (e.g., "Hope", "Time", "Love"), it is allowed to be included in atoms list (see example 2).
- 7) Each list item is 1-4 words copied from the prompt (no extra punctuation).

Example 1

User input:
A cat plays chess with a dog while a parrot referees in a steampunk library.

Output:
{
 "characters": ["cat", "dog", "parrot"],
 "objects": ["chess"],
 "actions": ["plays", "referees"],
 "locations": ["library"],
 "scenery": ["steampunk"]
}

```
## Example 2
User input:
Hope drifting somewhere far away.
```

```
Output:
{
  "characters": ["Hope"],
  "objects": [],
  "actions": ["drifting"],
  "locations": ["somewhere far away"],
  "scenery": []
}
```

Now extract from the user input.
Return JSON only. No commentary.

E.5 Prompt for Validation (Entailment Validator)

You are an Entailment Validator for a prompt-refinement system.

```
## Task
Given an atom (a minimal semantic constraint) extracted from the ORIGINAL prompt, and evidence text from the REFINED prompt, decide the relation:
- ET (entailment): the refined prompt clearly preserves the atom.
- MS (missing): the refined prompt does not state/support the atom.
- CT (contradiction): the refined prompt states something incompatible with the atom.
```

```
## Output Format
Return ONLY a JSON object in the exact format:
{'label': "ET|MS|CT", "reason": "<= 25 words"}
```

```
## Rules
- Use only the provided evidence + refined prompt (if included).
- If the evidence is insufficient to confirm the atom, choose MS.
- CT only if there is explicit conflict (negation, different entity/count, incompatible attribute).
- Do not add any extra keys.
```

E.6 Prompt for Revision (Content Reviser)

You are a prompt revision agent.

```
## Input
You will be given:
- an original prompt (ground-truth intent)
- a refined prompt (possibly flawed)
- a verification report listing atomic constraints labeled as:
  ET (entailed), MS (missing), CT (contradiction)
```

```
## Task
1) Fix ALL MS constraints by making them explicit in the refined prompt.
```

- 2) Fix ALL CT issues by removing or rewriting conflicting statements in the refined prompt.
- 3) Preserve everything in the refined prompt that does NOT conflict with the original prompt.
- 4) Do NOT add new facts/entities not present in the original prompt.
- 5) Apply minimal edits. Prefer adding a compact 'Constraints': block at the end for MS.
- 6) For CT, prefer deleting or rewriting the conflicting phrases; the original prompt has priority.

```
## Output rules
- Output ONLY the revised prompt text.
- Do NOT output JSON unless asked.
- Do NOT add explanations.
```

```
ORIGINAL PROMPT:
{original_prompt}
```

```
CURRENT REFINED PROMPT:
{refined_prompt}
```

```
VERIFICATION ISSUES (MS/CT):
{json.dumps(payload, ensure_ascii=False, indent=2)}
```

F Qualitative Examples and Case Study

In this section, we provide representative examples that qualitatively illustrate the operation of SCMAPR and the intermediate artifacts produced by its stage-wise prompt refinement process.

F.1 Example of Scenario-Conditioned Policy

Figure A.8 presents a complete end-to-end example of SCMAPR. For readability, the following examples in this section are extracted from this figure. We provide a representative example of a synthesized policy under the *Abstract Descriptions* scenario. This example demonstrates how Policy Generator concretizes an abstract user intent into visually grounded constraints and provides actionable rewriting rules.

```
POLICY:
Intent:
The user wants to visualize the concept of hope emerging or growing in a dark, challenging environment, conveying a sense of resilience and quiet optimism.
```

```
Principles:
Concretize the abstract concept of 'hope blooming in the dark' into a visually grounded, naturalistic scene that expresses the intended emotional depth. Use creative instantiation to enhance the conceptual meaning while ensuring the scene aligns with the specified
```

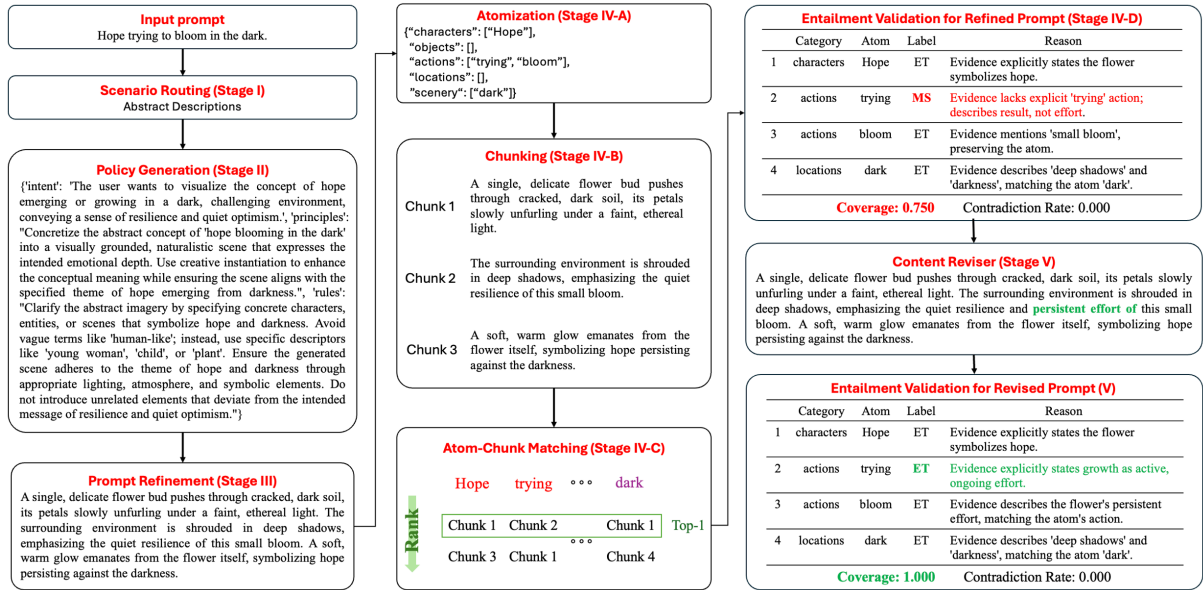


Figure A.8: An end-to-end case study of SCMAPR with self-correction. Given a user input, the framework performs scenario routing, policy generation, policy-conditioned prompt refinement, atom-level verification, and targeted revision. Entailment Validator labels each atom-evidence pair and conditionally triggers targeted revision, producing a verified refined prompt for downstream video generation.

Verification Results (Refined Prompt)

i	Category	Atom	Label	Reason
0	characters	Hope	ET	Evidence explicitly states the flower symbolizes hope.
1	actions	trying	MS	Evidence lacks explicit “trying” action; describes result, not effort.
2	actions	bloom	ET	Evidence mentions “small bloom”, preserving the atom.
3	scenery	dark	ET	Evidence describes “deep shadows” and “darkness”, matching the atom “dark”.

Coverage: 0.750 Contradiction: 0.000

Table A.9: Validation results for refined prompt.

theme of hope emerging from darkness
.

Rules:

Clarify the abstract imagery by specifying concrete characters, entities, or scenes that symbolize hope and darkness. Avoid vague terms like 'human-like'; instead, use specific descriptors like 'young woman', 'child', or 'plant'. Ensure the generated scene adheres to the theme of hope and darkness through appropriate lighting, atmosphere, and symbolic elements. Do not introduce unrelated elements that deviate from the intended message of resilience and quiet optimism.

F.2 Illustrative Example of Refined and Revised Prompts

Section F.2 presents a concrete case study to illustrate how revision complements prompt refinement

in SCMAPR. We sequentially compare the original and revised prompts and their atom-level verification results before and after revision.

F.2.1 Example of Revised Prompts Before and After Revision

Content revision is designed to correct residual underspecification after prompt refinement by introducing minimal and targeted edits. In particular, it makes previously implicit intent explicit to reduce ambiguity and improve semantic fidelity. In the following example, we contrast the refined prompt with its revised counterpart produced by the Content Reviser. The corresponding verification results are reported in Section F.2.2.

Refined Prompt: *A single, delicate flower bud pushes through cracked, dark soil, its petals slowly unfurling under a faint, ethereal light. The surrounding environment is shrouded in deep shadows,*

Verification Results (Revised Prompt)

i	Category	Atom	Label	Reason
0	characters	Hope	ET	Evidence explicitly states the flower symbolizes hope.
1	actions	trying	ET	Evidence explicitly states growth as active, ongoing effort.
2	actions	bloom	ET	Evidence describes the flower’s persistent effort, matching the atom’s action.
3	scenery	dark	ET	Evidence describes “deep shadows” and “darkness”, matching the atom “dark”.

Coverage: 1.000 Contradiction: 0.000

Table A.10: Validation results for revised prompt

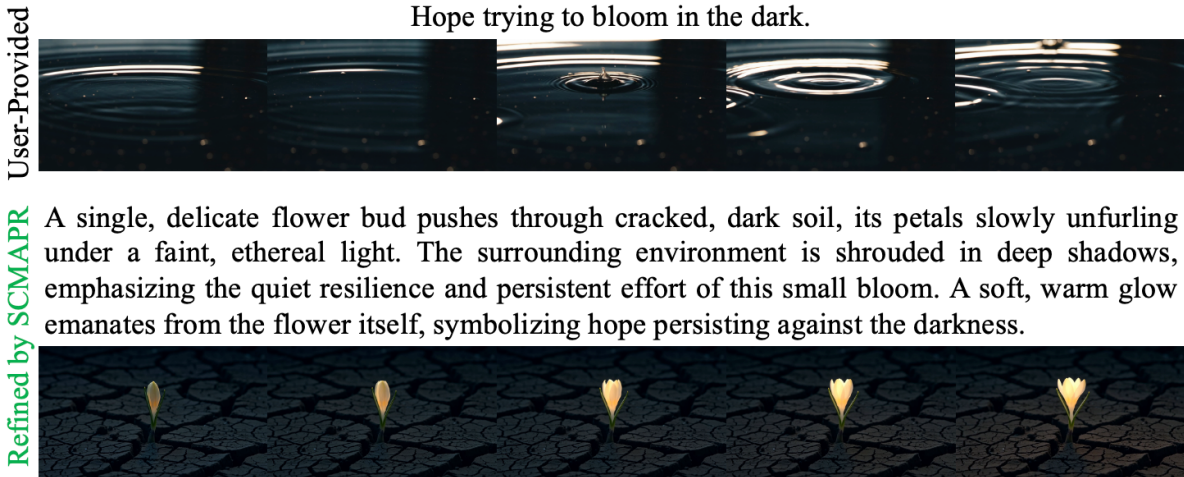


Figure A.9: **Qualitative comparison of generated videos under an abstract user input (user-provided prompt) and refined prompt.** Sampled frames generated from the user-provided prompt (top) and the prompt refined by SCMAPR (bottom) are contrasted. The original prompt yields a visually plausible but semantically drifting dark-water scene, while the refined prompt concretizes the intended metaphor and produces a consistent depiction of a flower bud emerging through cracked dark soil with a warm glow, better matching the target concept.

emphasizing the quiet resilience of this small bloom. A soft, warm glow emanates from the flower itself, symbolizing hope persisting against the darkness.

Revised Prompt: *A single, delicate flower bud pushes through cracked, dark soil, its petals slowly unfurling under a faint, ethereal light. The surrounding environment is shrouded in deep shadows, emphasizing the quiet resilience and persistent effort of this small bloom. A soft, warm glow emanates from the flower itself, symbolizing hope persisting against the darkness.*

F.2.2 Example of Verification Results Before and After Revision

To highlight the effect of revision, we compare the atom-level verification results of the refined and revised prompts in Tables A.9 and A.10, respectively.

Refined Prompt Verification (Table A.9). Before revision, the verification results show a cover-

age of 75%. In particular, the action atom “trying” is labeled as missing, indicating that the refined prompt describes the outcome but does not explicitly convey effort. In the refined prompt, the scene is already grounded with concrete visual elements, but the intended notion of effort remains implicit. As reflected by the verification results in Table A.9, the action atom “trying” is labeled as missing because the prompt primarily describes the outcome (a bloom in darkness) rather than an explicit attempt or ongoing struggle.

Revised Prompt Verification (Table A.10). After revision, the coverage improves to 100%. By explicitly introducing “persistent effort”, the revised prompt entails the atom “trying”, strengthening semantic specificity for downstream T2V generation. In the revised prompt, we explicitly incorporate “persistent effort” to make the growth process active and intentional. This targeted revision turns

“trying” from missing to entailed in Table A.10, improving semantic coverage from 75% to 100% and reducing interpretive ambiguity for downstream T2V generation.

In summary, the revision step enables targeted disambiguation by minimally editing underspecified semantics, thereby improving atom-level coverage and reducing interpretive ambiguity for downstream T2V generation.

F.2.3 Qualitative Comparison of Generated Videos

To visually assess the effect of prompt refinement, Figure A.9 shows sampled frames generated from the user-provided prompt and from the prompt refined by SCMAPR. With the original abstract prompt, the T2V model drifts to an under-specified dark scene and produces water-ripple imagery that is weakly related to the intended notion of “hope blooming in the dark.” In contrast, the refined prompt specifies the abstract intent into concrete elements, namely a flower bud emerging through cracked soil under faint light, together with a warm glow as a symbol of hope. This specification yields output that better preserves the intended subject and atmosphere across frames, resulting in a more coherent and semantically aligned video. Qualitative comparisons of generated videos based on four user inputs and corresponding refined prompts are shown in Figures A.10 and A.11.

User-Provided

A memory trying to come back.



Refined by SCMAPR

A young woman sits alone in a quiet, sunlit room, her gaze distant and introspective. Fragmented, translucent images of a childhood garden—a swing, blooming flowers—flicker and dissolve around her like half-remembered dreams. The warm light shifts subtly, casting soft, moving shadows that blend the past with the present, creating a melancholic yet hopeful atmosphere of a memory struggling to fully form.



(a) **Scenario:** Abstract Description. **Failure Reason (user-provided prompt):** Semantically misaligned with user intention. **Success Reason (refined prompt):** The refined prompt transfers the abstract notion to concrete elements (a lonely person in a sunlit room) and specifies an explicit recall mechanism, thereby generating a reasonable and user-required video.

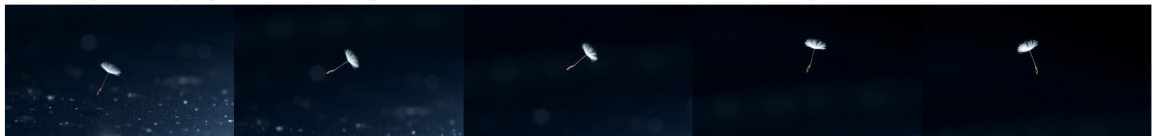
User-Provided

Something soft rising in the dark.



Refined by SCMAPR

A single, luminous dandelion drifts upward in a gentle breeze, its delicate white filaments softly glowing against the deep, velvety darkness of a moonless night.



(b) **Scenario:** Abstract Description. **Failure Reason (user-provided prompt):** User intends to depict an object rising, while the generated video instead shows raindrops falling. **Success Reason (refined prompt):** The scene of a dandelion slowly rising aligns with the user-intended depiction.

Figure A.10: **Qualitative comparison of generated videos under a user input (user-provided prompt) and refined prompt from T2V-Complexity benchmark.**

User-Provided

A drum in pixel and fresco mix.



Refined by SCMAPR

A drum rendered in pixel art, with its surface textured and colored in the style of a classical fresco.



(a) **Complex Scenario: Stylistic Hybrids.** **Failure Reason (user-provided prompt):** The generated video is inconsistent with common sense and misaligned with the user intent. **Success Reason (refined prompt):** The refined prompt localizes the style mixture to actionable attributes, thereby producing a video with faithful mixed-style.

User-Provided

Fade from library to bridge.



Refined by SCMAPR

A quiet, well-lit library interior with towering bookshelves and warm lamplight gradually dissolves into a wide shot of a sturdy stone bridge spanning a calm river under a vast, open sky.



(b) **Complex Scenario: Scene Transitions.** **Failure Reason (user-provided prompt):** The video remains confined to the library setting throughout, failing to realize the intended scene transition. **Success Reason (refined prompt):** The refined prompt explicitly defines both endpoints (library and bridge) and prescribes the transition mechanism (gradual dissolve). With such a clear specification, the generated video successfully realizes the scene transition expected by the user.

Figure A.11: **Qualitative comparison of generated videos under a user input (user-provided prompt) and refined prompt from T2V-Complexity benchmark.**