

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 CSAPR: COMPLEX-SCENARIO-AWARE PROMPT RE- FINEMENT FOR TEXT-TO-VIDEO GENERATION

Anonymous authors

Paper under double-blind review

## ABSTRACT

Recent years have witnessed rapid progress of diffusion models, which significantly advance the development of Text-to-Video (T2V) generation. Compared to Text-to-Image (T2I) generation, T2V models encounter additional challenges, including temporal consistency, motion coherence, and adherence to physical constraints across frames. To address these challenges, we propose a novel two-stage framework, i.e., Complex-Scenario-Aware Prompt Refinement (CSAPR), to improve prompt the quality for T2V generation. CSAPR consists of two stages, i.e., prompt refinement and prompt verification. In the prompt refinement stage, CSAPR classifies user prompts into one of eight representative categories and applies targeted rewriting strategies guided by predefined meta prompts. In the prompt verification stage, CSAPR aligns semantic atoms from the original prompt with decomposed chunks of the refined prompt, ensuring that the refined prompt faithfully preserves the intended semantics while avoiding inconsistencies. Extensive experiments on three benchmarks, i.e., VBench, EvalCrafter, and T2V-CompBench, demonstrate that CSAPR significantly improves alignment with user intent and video generation quality in complex scenarios (up to 1.40% in terms of average score).

## 1 INTRODUCTION

The rapid advancement of diffusion models (Peebles & Xie, 2023; Rombach et al., 2022) has revolutionized 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. Amid these advances, Text-to-Video (T2V) generation has emerged as a particularly compelling direction. Compared to Text-to-Image (T2I) generation, T2V demands visually realistic frames together with temporal coherence, motion dynamics, and adherence to physical and causal constraints, thereby posing a substantially more challenging yet impactful research problem.

Recent works (Hao et al., 2023; Zhan et al., 2024b;a) have demonstrated that the prompts optimized by Large Language Models (LLMs) yield high-quality and user-aligned outputs when fed into diffusion models. This improvement arises while user-provided prompts tend to be concise and may omit the detailed descriptions favored by AIGC models. Specifically, high-quality prompts are expected to provide precise character and scene descriptions, follow specific expression patterns and incorporate domain-specific terminology for stylistic expression (Parsons, 2022; Witteveen & Andrews, 2022; Brade et al., 2023; Zhan et al., 2024a).

However, existing prompt refinement approaches are generally designed for 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 these strategies are often sufficient for producing a single coherent image, they fall short when applied to T2V generation, which introduces substantially greater complexity. In video generation tasks, models are required to simultaneously guarantee temporal consistency, ensure motion coherence, capture causal dependencies, and comply with physical laws across frames. Although RAPO (Gao et al., 2025) attempts to enrich textual

descriptions by constructing a relation graph to retrieve terms relevant to user prompts, it primarily focuses on inter-object relations while overlooking challenges distinctive to video generation, such as abstract descriptions or temporal consistency. More importantly, RAPO relies excessively on relevance-based prompt retrieval and augmentation, which cannot provide a deeper understanding of the motivations behind prompt rewriting or explicit guidance for refinement in complex scenarios.

To address these challenges, we summarize seven representative complex scenarios that pose significant difficulties for diffusion models in generating videos that align with user expectations. Since each scenario introduces unique challenges for T2V generation, distinct strategies are required to address them systematically. Motivated by this insight, we propose a novel two-stage framework, i.e., Complex-Scenario-Aware Prompt Refinement (CSAPR).

CSAPR is designed to enable the prompt refinement model to understand the underlying reasons why a generation request may be difficult for a T2V model to fulfill, and to employ targeted rewriting strategies accordingly. To achieve this, CSAPR adopts a two-stage framework that includes prompt refinement and prompt verification. In the refinement stage, CSAPR classifies the user input into one of the seven challenging scenarios or identifies it as a non-challenging case. According to the classification result, a pre-defined rewriting guideline is selected and employed as a meta-prompt to guide the rewriting process. In the verification stage, CSAPR extracts the original prompt into atoms and decomposes the rewritten prompt into multiple chunks. A semantic alignment process is then performed to verify whether the rewritten chunks adequately preserve and reflect the meaning of the original atoms. If verification is successful, the refined prompt is delivered to the T2V model. Otherwise, the rewriting model is provided explicit feedback about missing or conflicting semantic elements and instructed to regenerate the refined prompt. The major contributions are summarized as follows:

- We summarize seven challenging scenarios in T2V generation tasks and propose targeted improvement strategies for each scenario. To our knowledge, this is the first work that investigates prompt refinement for video generation in complex scenarios.
- Our prompt rewriting strategy is grounded in understanding the challenges faced by T2V models and is explicitly guided by well-defined objectives throughout the rewriting process.
- We additionally design a prompt verification method to ensure that the rewritten prompts neither omits any user-specified elements nor introduces conflicts with the user prompt.
- Extensive evaluations on VBench, EvalCrafter, and T2V-CompBench demonstrate the effectiveness of our prompt refinement approach for T2V generation tasks.

## 2 RELATED WORK

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

### 2.1 TEXT-TO-VIDEO GENERATION

With the rapid progress of diffusion transformers and large-scale generative models (Rombach et al., 2022; Peebles & Xie, 2023), T2V generation (Singer et al., 2023; Chen et al., 2024a) has emerged as a pivotal task in content creation. Building on this trend, recent research has advanced T2V models along multiple directions. Architectural innovations have introduced scalable designs, such as expert transformers (Yang et al., 2025) and linear-complexity attention modules (Wang et al., 2025b), which improve both efficiency and model capacity. Training-free and plug-and-play inference approaches further enhance motion dynamics and spatial fidelity without additional training by exploiting cross-model integration and attention map analysis (Bu et al., 2025; Zhang et al., 2025b; Jagpal et al., 2025). Precise control over entity appearance and interactions has also been achieved through structured captions, instance-aware modeling, and LoRA-based customization (Feng et al., 2025; Fan et al., 2025; Huang et al., 2025a). Beyond visual quality, LLM-guided reasoning and external knowledge retrieval are employed to improve adherence to physical laws and factual accuracy (Xue et al., 2025; Yuan et al., 2025). Video editing frameworks exploit spatial-temporal guidance and attention control to extend T2V models for precise video modification (Wang et al., 2025e). Although these methods have made significant progress in various specific directions, they focus on common scenarios, while video generation for complex scenarios remains a challenging task. For instance,

108 reasoning about complex multi-entity interactions and maintaining coherent event-level narratives  
 109 are still open problems, especially in scenes requiring abstract semantics or long-range temporal  
 110 dependencies.

111

## 112 2.2 PROMPT REFINEMENT

113

114 Prompt refinement aims to automatically optimize user-provided natural language prompts into en-  
 115 hanced formulations that better align with the preferences of diffusion models. Early studies primar-  
 116 ily infer user preferences or rewriting capabilities to guide prompt refinement. Prompt Refinement  
 117 with Image Pivot (PRIP) (Zhan et al., 2024a) encodes visual preferences from linguistic prompts into  
 118 latent representations, which are then decoded into refined text prompt. Capability-Aware Prompt  
 119 Reformulation (CAPR) (Zhan et al., 2024a) dynamically adjusts rewriting strategies based on user  
 120 capability and introduces configurable features for fine-grained control. Although effective, these  
 121 approaches explicitly rely on user click logs, system-generation logs and other interaction data,  
 122 which assume large-scale user engagement with text-to-image or text-to-video models and require  
 123 explicit user consent for data collection. More recent studies incorporate retrieval-augmented gener-  
 124 ation (RAG) techniques to enhance prompts by retrieving semantically relevant terms and enriching  
 125 the original descriptions. These methods either construct a dedicated prompt repository as an exter-  
 126 nal knowledge base (Sun et al., 2024) or build a relation graph from training data (Gao et al., 2025).  
 127 At inference time, they retrieve the most relevant exemplar prompts and integrate them with user  
 128 input to compose a coherent, optimized reformulation. However, these RAG-based methods merely  
 129 find high-similarity descriptions to enrich original prompts, without reasoning about the causes for  
 130 rewriting or specifying explicit reformulation goals, limiting their ability to address complex gener-  
 131 ation scenarios.

132

## 133 3 COMPLEX-SCENARIO-AWARE PROMPT REFINEMENT

134

135 In this section, we first present the background and motivation. Then, we explain the scenario  
 136 classification-based prompt refinement. Afterward, we propose our four-stage prompt verification.

137

### 138 3.1 BACKGROUND AND MOTIVATION

139

140 In real-world scenarios, T2V models may receive highly challenging generation requests, such as  
 141 synthesizing abstract concepts or producing videos with intricate inter-object relationships. Un-  
 142 like text-to-image tasks, T2V generation further entails challenges of maintaining temporal coher-  
 143 ence and ensuring logical consistency across frames. For example, a prompt describing “a flower  
 144 blooming” requires modeling fine-grained morphological changes over time, while “multi-person  
 145 interactions” demands precise control over dynamic relationships between characters. Such scenar-  
 146 os are highly susceptible to bring up prompt ambiguity, leading to unexpected generated videos,  
 147 which cannot precisely meet the original prompt intention. With our excellent rewriting strategy,  
 148 prompt refinement transforms abstract or complex user intents into explicit structured descriptions  
 149 that specify action sequences, spatial relationships, and physical constraints. As shown in Figure 1,  
 150 the proposed approach CSAPR consists of two complementary stages, i.e., prompt refinement and  
 151 prompt verification. CSAPR aims to unlock the full potential of T2V generation systems through  
 152 precise scene descriptions and detailed depictions of event dynamics.

153

### 154 3.2 REFINEMENT STAGE: SCENARIO CLASSIFICATION-BASED PROMPT REFINEMENT

155

156 In this section, we propose our scenario classification-based prompt refinement, including scenario  
 157 summarization, refinement strategies for each scenario, and scenario classification.

158

#### 159 3.2.1 SCENARIO SUMMARIZATION

160

161 In order to systematically identify representative challenges in T2V generation, we conduct a com-  
 162 prehensive review of recent benchmarks, surveys, and task-specific studies (Furuta et al., 2024; Liao  
 163 et al., 2024; Mago et al., 2025; Sun et al., 2025; Wang et al., 2025c). In addition, we carry out an  
 164 comprehensive examination of existing open source T2V models, e.g., Wan (Wang et al., 2025a),  
 165 OpenSora (Zheng et al., 2024), CogVideo (Hong et al., 2022), and proprietary T2V systems, e.g.,

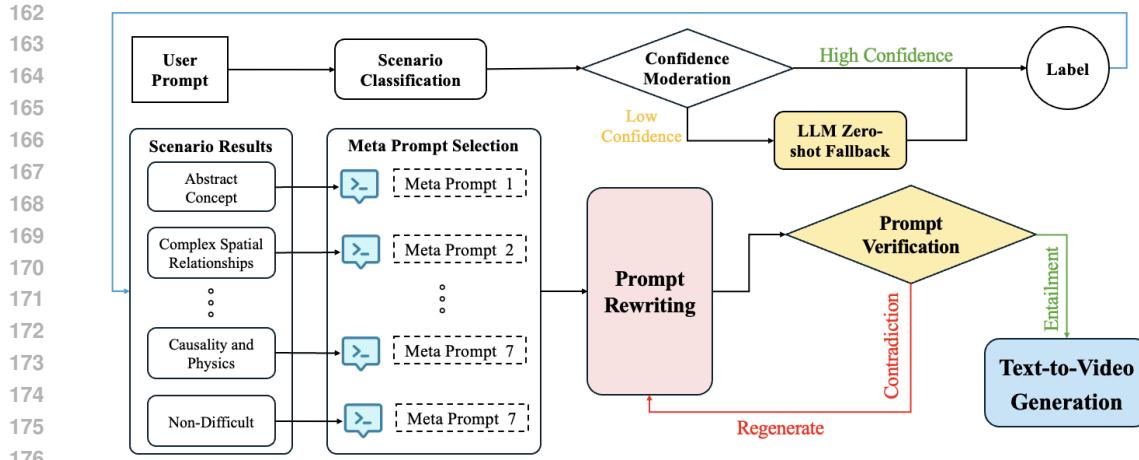


Figure 1: **Pipeline of Complex-Scenario-Aware Prompt Refinement (CSAPR).** CSAPR comprises two primary stages: prompt refinement and prompt verification. In the refinement stage, the input prompt is rewritten using a meta-prompt determined by its scenario classification. In the verification stage, the refined prompt is examined to ensure that it preserves the complete semantic content of the original prompt.

Veo3 (Google DeepMind, 2025), OpenAI Sora (OpenAI, 2024), Runway Gen-2 (Jennings & Runway, 2023).

From the review and the examination, we summarize seven representative categories of scenarios that are widely acknowledged as challenging for current T2V models. Abstract descriptions rely on metaphors or symbolism that require translation into concrete, visually interpretable content. Complex spatial relationships involve precise positions and orientations among multiple entities, where models often misrepresent layouts or occlusions. Multi-element scenes contain numerous characters, objects, and actions, demanding structured descriptions to avoid omissions or disorder. Fine-grained details emphasize subtle attributes such as facial expressions or textures, which models frequently blur or distort. Temporal consistency requires coherent progression of actions across frames, yet T2V models generally produce discontinuities or sudden jumps. Stylistic hybrids call for the integration of heterogeneous visual styles, which can result in incoherent or conflicting appearances. Lastly, causality and physics demand accurate cause–effect reasoning and physical plausibility, where T2V models typically fail to simulate realistic interactions or transformations. See representative examples in Appendix A.

### 3.2.2 SCENARIO-SPECIFIC REFINEMENT STRATEGIES

In order to facilitate fine-grained prompt refinement strategies for each challenging scenario, we design scenario-specific meta-prompts that guide a rewriting model to generate more precise scene descriptions and richer representations of event dynamics compared with the original prompts, with detailed prompts provided in Appendix E. In addition, we define meta-prompts for non-challenging scenarios to encourage the rewriting model to explicitly specify spatial relationships among objects and temporal dependencies among events.

In abstract descriptions, the rewriting model is guided to translate metaphors and symbolic language into concrete, visually interpretable content while preserving the intended artistic or thematic meaning. For complex spatial relationships, the rewriting model is encouraged to convey precise positions, distances, and orientations among multiple entities to avoid misrepresentations. Multi-element scenes require logically structured descriptions that comprehensively account for all characters, objects, and actions to prevent omissions or disorder. Fine-grained details are emphasized by highlighting subtle visual attributes and incorporating cinematic cues that capture textures, expressions, and other intricate elements. Temporal consistency is maintained by presenting actions in coherent sequences and smooth motion progression, while stylistic hybrids are addressed by harmonizing multiple visual styles, including color palettes, lighting, and composition. Causality and

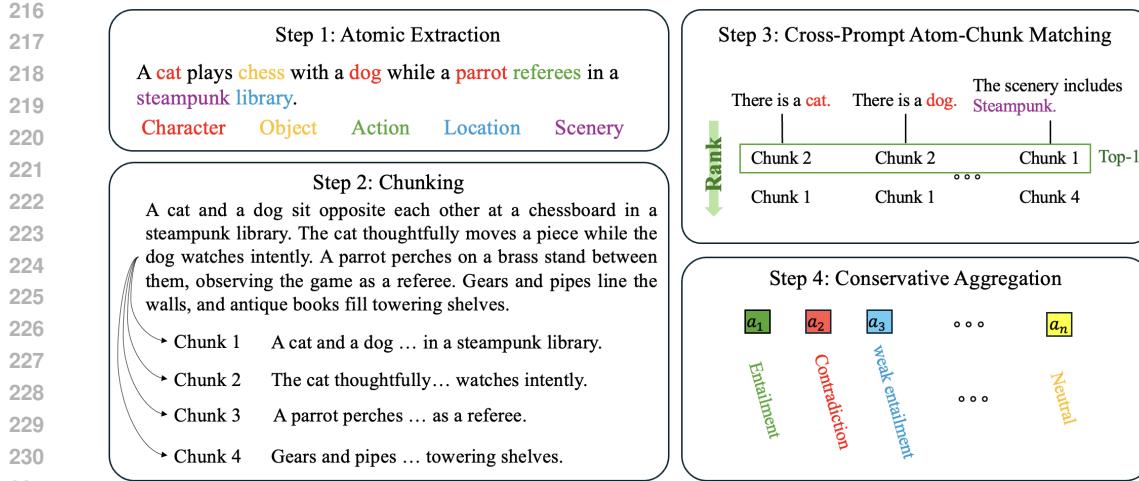


Figure 2: **Pipeline of Refined Prompt Verification via Atomic Extraction and Textual Entailment.** Our goal is to check whether the refined prompt  $P_{\text{rew}}$  faithfully preserves the original intent in  $P_{\text{in}}$ . The process consists of four main stages: atomic extraction, semantic chunking, atom–chunk alignment, and conservative aggregation.

physical plausibility are reinforced by explicitly depicting cause–effect relationships, physical interactions, and stepwise morphological changes. For non-challenging scenarios, prompts encourage light refinement to improve clarity while preserving the original intent.

### 3.2.3 SCENARIO CLASSIFICATION

After receiving a user prompt  $P_{\text{in}}$ , the scenario classifier  $f$  performs an eight-way classification with seven difficult scenarios and a non-difficult class. The classifier classifies the original prompt to the most suitable scenario label  $\hat{y}$  and estimates the confidence of this classification. To ensure the reliability of the scenario classification, we introduce three confidence metrics, i.e., *Maximum Softmax Probability* (MSP), *Margin*, and *Entropy*. Given a set of class probabilities  $\{p_i\}_{i=1}^C$  over candidate labels  $C$ , these metrics can be computed as follows:

$$\text{MSP} = \max_i p_i; \quad \text{Margin} = p_{(1)} - p_{(2)}; \quad \text{Entropy} = - \sum_{i=1}^C p_i \log p_i, \quad (1)$$

where  $p_{(1)}$  and  $p_{(2)}$  denote the top-1 and top-2 probabilities, respectively. Intuitively, a confident prediction is characterized by a high MSP. A large Margin indicates clear separation between the top predictions. A low Entropy reflects a peaked probability distribution. A classification result is considered uncertain if MSP is below a predefined confidence threshold, Margin is small, or Entropy is high.

Once a classification is marked as uncertain, the scenario label is reassigned through a more reliable large language model (e.g., ChatGPT (Achiam et al., 2023), Gemini (Team et al., 2023), or DeepSeek (Liu et al., 2024a)), which performs zero-shot classification based on predefined scenario definitions as a fallback strategy. The final predicted label  $\hat{y}$  corresponds to a *scenario-specific meta-prompt*, which is applied to guide prompt refinement. This meta-prompt is then combined with the original user prompt and fed into the prompt refinement model to produce a rewritten prompt  $P_{\text{rew}}$ . Then,  $P_{\text{rew}}$  are required to undergo a dedicated prompt verification stage (see Section 3.3). Only validated prompts  $P_{\text{valid}}$  are passed to the downstream T2V generation model, while invalid prompts  $P_{\text{invalid}}$  are discarded. If a rewritten prompt is judged invalid, the refinement model will regenerate an improved prompt according to the verification feedback. For instance, if the rewritten prompt omits semantic information from the original user prompt, an additional instruction reminding the model to preserve this information will be concatenated to the end of meta-prompt before the user prompt.

270  
271

## 3.3 PROMPT VERIFICATION

272  
273  
274  
275  
276  
277  
278  
279  
280

In this section, we propose a structured verification framework to rigorously assess whether an expanded T2V generation prompt faithfully retains the semantic essence of its original concise input. Expanded prompts generally incorporate stylistic attributes and narrative embellishments, therefore reliable verification is critical to ensure the quality of T2V generation. As shown in Figure 2, to address this challenge, our method decomposes the process into four well-defined stages: (1) atomic extraction: isolating fundamental entities and actions; (2) semantic alignment: ensuring coverage and consistency of critical elements; (3) dual-model textual entailment: leveraging complementary inference models to evaluate semantic fidelity; and (4) conservative aggregation: synthesizing multi-stage evidence to produce a robust verification judgment.

281  
282

## 3.3.1 ATOMIC EXTRACTION FROM ORIGINAL PROMPT

283  
284  
285

Given an original prompt  $P_{\text{user}}$ , we define an atomic representation as a minimal semantic tuple:

$$\mathcal{A}(P_{\text{orig}}) = \langle \text{characters}, \text{objects}, \text{actions}, \text{locations}, \text{scenery} \rangle. \quad (2)$$

286  
287  
288  
289  
290  
291  
292  
293  
294

Each field in  $\mathcal{A}$  is a list that can contain zero, one, or more atomic elements  $a_i$ , where each element  $a_i$  corresponds to a semantically indivisible unit in that category. For example, given the prompt “*A cat plays chess with a dog while a parrot referees in a steampunk library*”, the atomic extractor produces: {characters: [cat, dog, parrot], objects: [chess], actions: [plays, referees], locations: [library], scenery: [steampunk] }. After extracting atomic entities, we expand each atom into a minimal standalone sentence to facilitate similarity computation with the chunked segments of the rewritten prompt. For example, the atom cat is transformed into a simple sentence as “there is a cat”. To simplify processing, we discard the dictionary keys and retain the atomic elements themselves, forming a set that is formally represented as  $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$ .

295  
296

## 3.3.2 REFINED PROMPT CHUNKING

297  
298  
299

The refined prompt  $P_{\text{ref}}$  is segmented into non-overlapping chunks  $\mathcal{C} = \{c_1, \dots, c_m\}$ . One sentence is treated as an individual chunk unless its length is below a threshold. In that case, it is iteratively merged with the subsequent sentence until the combined length exceeds the threshold.

300  
301  
302  
303  
304  
305  
306  
307  
308

**The Necessity of Chunking and Atomization.** A naive strategy would be to directly compute the similarity between the embeddings of the original and refined prompts. However, this approach can only provide a coarse measure of semantic closeness and does not clearly determine whether the refined prompt omits critical information from the original. In addition, refined prompts are commonly longer and contain additional stylistic and contextual details, which tend to shift their embeddings away from those of short original prompts even when semantics are faithfully preserved. More importantly, similarity does not distinguish between inclusion and contradiction. For example, the atom “a girl in a red raincoat” and the expansion “a girl in a yellow raincoat running at night” can still exhibit a high embedding similarity, although they are semantically inconsistent.

309  
310

## 3.3.3 CROSS-PROMPT ATOM-CHUNK MATCHING

311  
312  
313  
314

We adopt a stricter entailment-based criterion rather than merely comparing the similarity between the original and refined prompts. Specifically, we evaluate whether the refined prompt entails each atomic constraint. Note that similarity is leveraged only to retrieve candidate text chunks, while the final verification is performed entirely by textual entailment models.

315  
316  
317  
318

Specifically, given a set of atoms  $\{a_i\}_{i=1}^m$  and a set of chunks  $\{c_j\}_{j=1}^n$ , we embed both atoms and chunks into a shared semantic space using an embedding model  $f_{\mathcal{E}}(\cdot)$ , and compute pairwise cosine similarities:

$$s_{ij} = \cos(f_{\mathcal{E}}(a_i), f_{\mathcal{E}}(c_j)).$$

320  
321

For each atom  $a_i$ , we typically select the most relevant chunk (i.e., Top-1 retrieval) from the collection:

$$c_i^* = \arg \max_{c \in \{c_j\}} s(f_{\mathcal{E}}(a_i), f_{\mathcal{E}}(c)), \quad (3)$$

322  
323

where  $c_i^*$  denotes the retrieved chunk with the highest similarity score.

324 Table 1: Quantitative comparisons on VBench. Boldfaced values indicate the best performance  
 325 among all methods, while underlined values indicate the second-best performance.  
 326

327 <b>Method</b>	328 <b>Average Score</b>	329 temporal flickering	330 imaging quality	331 object class	332 multiple objects	333 spatial relationship
Original Prompt	76.38%	97.02%	70.30%	92.09%	65.32%	57.14%
Open-Sora	<u>77.22%</u>	<u>97.58%</u>	70.76%	93.61%	65.84%	58.29%
RAPO	<u>79.11%</u>	<u>97.94%</u>	<u>71.43%</u>	<u>97.13%</u>	66.68%	<u>62.37%</u>
CSAPR	<b>79.78%</b>	<b>98.52%</b>	<b>71.89%</b>	<b>97.85%</b>	<b>67.43%</b>	<b>63.22%</b>

331 Table 2: Quantitative comparisons on EvalCrafter. CSAPR consistently achieves better results,  
 332 demonstrating significant lead on this benchmark.  
 333

334 <b>Method</b>	335 <b>Average Score</b>	336 motion quality	337 text-video alignment	338 visual quality	339 temporal consistency
Original Prompt	62.54	53.48	70.83	64.77	61.09
Open-Sora	63.01	53.61	71.05	65.41	61.98
RAPO	<u>63.98</u>	<u>53.64</u>	<u>74.33</u>	<u>65.93</u>	<u>62.02</u>
CSAPR	<b>65.38</b>	<b>54.12</b>	<b>76.41</b>	<b>66.24</b>	<b>64.75</b>

### 340 3.3.4 TEXTUAL ENTAILMENT VERIFICATION

341 To assess whether each atom is semantically preserved by its most relevant retrieved chunk, we employ a *textual entailment verification* module based on Natural Language Inference (NLI). Specifically, for each atom  $a_i$  and its top-1 retrieved chunk  $c_i^*$ , we employ XLM-RoBERTa (Conneau et al., 342 2020) as the multilingual NLI model to directly predict the semantic relation label from entailment (ENTAIL), neutral (NEUTRAL), or contradiction (CONTRAD).

343 Verification is evaluated through two primary metrics. Coverage (Cov) quantifies the proportion 344 of atoms that are entailed by their corresponding chunk  $\text{Cov} = \frac{|\{a_i: \text{ENTAIL}\}|}{|\mathcal{A}|}$ . Contradiction rate 345 measures the proportion of atoms contradicted by their chunk:  $\text{Contrad} = \frac{|\{a_i: \text{CONTRAD}\}|}{|\mathcal{A}|}$ . Coverage 346 reflects the semantic completeness of the expanded prompt, while contradiction rate highlights 347 conflicts introduced during prompt refinement. When coverage falls below a predefined threshold 348 or contradiction exceeds a threshold, the system triggers the prompt refinement model to regenerate 349 an improved prompt based on the verification feedback. Specifically, if the refined prompt omits 350 semantic content present in the original user prompt, an additional instruction reminding the model to 351 preserve this information will be appended to the end of the meta-prompt preceding the user prompt.  
 352

## 353 4 EXPERIMENTS

354 In this section, we present the experimental settings with three benchmarks. Then, we demonstrate 355 the main experimental results. Finally, we show an ablation study.  
 356

### 357 4.1 EXPERIMENTAL SETTINGS

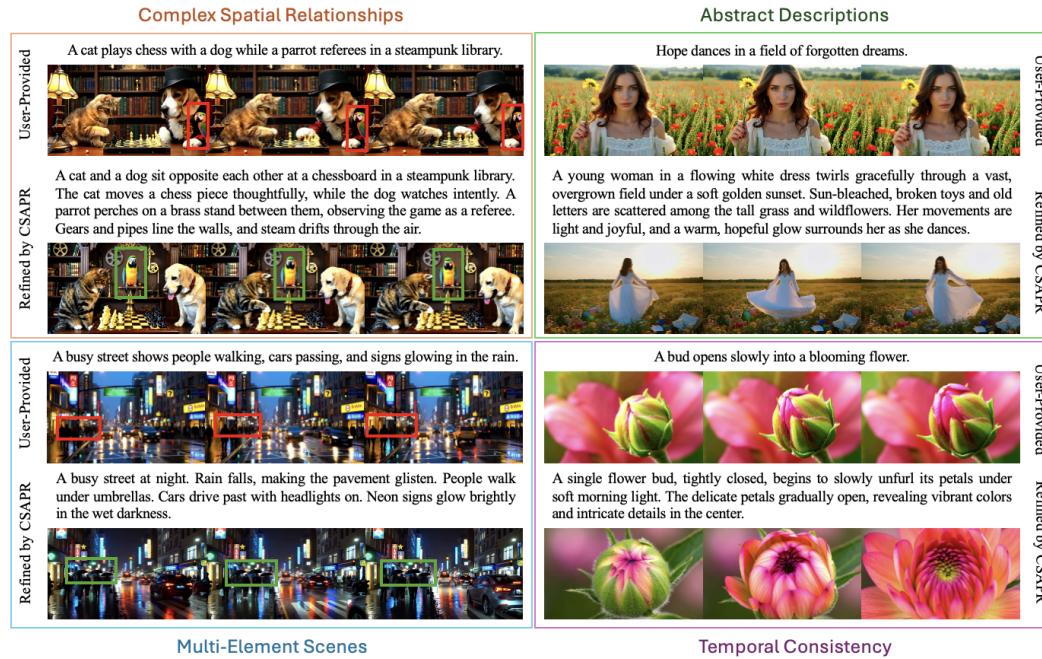
358 **Benchmarks:** We conduct evaluations on VBench (Huang et al., 2024), EvalCrafter (Liu et al., 359 2024c) and T2V-CompBench (Sun et al., 2025), three state-of-the-art benchmarks to evaluate the 360 quality of T2V generation. VBench provides a systematic evaluation protocol for comprehensive 361 assessment of visual quality, temporal consistency, and content fidelity. EvalCrafter offers a broad suite 362 of metrics to quantify performance across multiple aspects of video generation. T2V-CompBench is 363 a benchmark specifically designed for compositional T2V scenarios.  
 364

365 **Baselines:** We compare three representative prompt refinement approaches in the field of T2V, 366 including: the original prompts, prompt refiner from Open-Sora (Zheng et al., 2024) and Retrieval- 367 Augmented Prompt Optimization (RAPO) (Gao et al., 2025).  
 368

369 **Implementation:** In the prompt refinement setting, we adopt Wan (Wang et al., 2025a) as the T2V 370 backbone for all experiments. For scenario classification, DeBERTa-v3-large (He et al., 2023) is ex- 371 ploited as the primary model, while DeepSeek-V3 (DeepSeek) (Liu et al., 2024a) serves as a zero- 372 shot fallback. Then, we query a large instruction-tuned LLM, i.e., DeepSeek, to generate refined 373 prompts based on scenario-specific meta-prompts (see Appendix E). In the settings of prompt verifi-  
 374

378 Table 3: Quantitative comparisons on T2V-CompBench. CSAPR achieves the highest average score.  
379

380 <b>Method</b>	381 <b>Average Score</b>	382 consistent attribute	383 dynamic attribute	384 action binding	385 motion binding
Original Prompt	0.412	0.628	0.254	0.478	0.290
Open-Sora	0.433	0.672	0.269	0.493	0.298
RAPO	<u>0.470</u>	<u>0.682</u>	<u>0.270</u>	<u>0.612</u>	<u>0.317</u>
CSAPR	<b>0.501</b>	<b>0.724</b>	<b>0.289</b>	<b>0.641</b>	<b>0.352</b>

407 Figure 3: Comparisons of videos generated using Wan (Wang et al., 2025a) conditioned on user-  
408 provided prompts and refined prompts from CSAPR.  
409410  
411 cation, we employ BGE-M3 (Chen et al., 2024b) as the embedding model for atom-chunk matching  
412 in prompt verification. On the textual entailment task, XLM-RoBERTa-large-XNLI (Conneau et al.,  
413 2020) serves as the textual entailment model for atom-level entailment checking.414  
415 

## 4.2 MAIN RESULTS

416  
417 As shown in Tables 1, 2 and 3, we present a comprehensive quantitative evaluation of our proposed  
418 approach, i.e., CSAPR, against baselines across three widely used benchmarks. It can be observed  
419 that CSAPR consistently achieves the best overall performance, demonstrating its effectiveness in  
420 T2V generation. Specifically, CSAPR attains the highest Average Score of 79.78% on VBench,  
421 outperforming RAPO by 0.67%, Open-Sora by 2.56% and the Original Prompt by 3.40%. Similar  
422 trends can also be observed on EvalCrafter, where CSAPR achieves the highest Final Sum Score of  
423 260.21, which significantly surpasses other approaches. Notably, it shows marked improvements in  
424 text-video alignment and temporal consistency, underscoring its strength in semantic fidelity and dy-  
425 namic coherence. When it comes to T2V-CompBench focuses on compositional reasoning, CSAPR  
426 achieves the best results in consistent attribute binding, dynamic attribute binding, action binding  
427 and motion binding, demonstrating strong performance in fine-grained attribute and motion model-  
428 ing. In particular, CSAPR surpasses the RAPO, Open-Sora, and original prompt by 0.055, 0.092,  
429 and 0.113 on T2V-CompBench when it comes to average score.430  
431 To provide an intuitive demonstration on the advantages of CSAPR, Figure 3 highlights the effec-  
432 tiveness of CSAPR across four categories of complex scenarios: complex spatial relations, abstract  
433 descriptions, multi-entity scenes and temporal consistency. In the case of complex spatial relations,  
434 CSAPR achieves two notable improvements over videos generated directly from the user prompt:

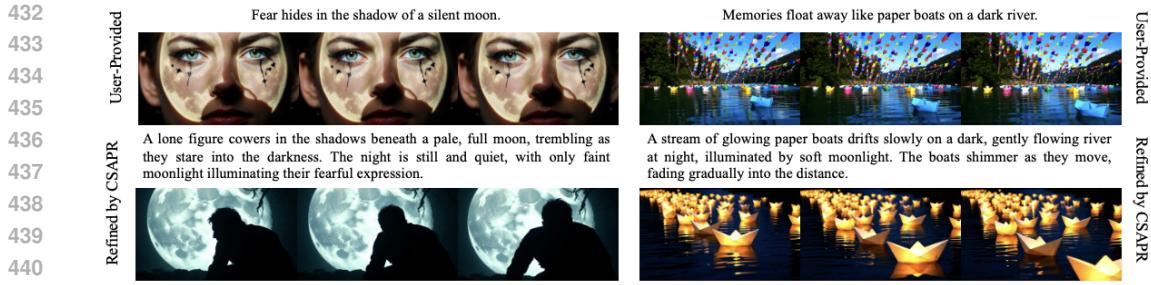


Figure 4: Examples of hallucination elimination after prompt refined by CSAPR.

(1) the parrot is correctly placed at the center (green box) rather than at the edge (red box); and (2) the library is rendered with cyberpunk elements instead of being depicted as a regular library. For abstract descriptions, CSAPR produces richer actions and more vivid visual dynamics. In multi-entity scenes, CSAPR meticulously introduces umbrellas for pedestrians in rainy weather (green box). Finally, under temporal consistency, while the baseline fails to present a fully blossomed flower, CSAPR generates the complete blooming process, resulting in a fully blossomed flower video.

**Hallucination Elimination.** Figure 4 presents examples where simple prompt descriptions lead to hallucinations. In the left panel, the original abstract prompt 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 CSAPR provides a concrete depiction aligned with the abstract prompt. Therefore, the T2V model can well understand the desired content and then eliminate hallucinations with the refined prompts of CSAPR. 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 CSAPR conveys a coherent visual narrative that captures the fading of memories and the lingering attachment to the past.

### 4.3 ABLATION STUDY

To verify the effectiveness of CSAPR, we conduct ablation experiments on the VBench benchmark to examine four key modules in CSAPR. Table 4 presents the ablation results on VBench. The Complete CSAPR achieves the best performance. Either removing the LLM zero-shot fallback or prompt verification leads to minor performance drops (0.87 and 1.05 higher score, respectively). In contrast, discarding the scenario classification (1.80 lower score) or prompt refinement modules (2.69 lower score) results in relatively larger degradations, highlighting their critical roles in CSAPR.

## 5 CONCLUSION

In this paper, we propose a novel prompt refinement approach, i.e., Complex-Scenario-Aware Prompt Refinement (CSAPR), designed for complex T2V generation tasks. CSAPR consists of two key stages: prompt refinement and prompt verification. The prompt refinement stage involves complex scene classification, confidence estimation and LLM-based zero-shot fallback. Following this, the prompt verification stage comprises atomic extraction of the original prompt, new prompt chunking, cross-prompt atom-chunk matching and conservative aggregation to ensure the coherence and completeness. Extensive experiments demonstrate that CSAPR consistently outperforms baseline approaches (up to 1.40% higher in terms of average score).

486 REPRODUCIBILITY STATEMENT  
487488 To ensure the reproducibility of our work, we have provided comprehensive details about the ex-  
489 perimental setup in Section 4, including datasets used, baseline methods, implementation details and  
490 evaluation metrics. More details about the implementation can be found in the Appendix. All code,  
491 models, and configuration files required to replicate our results are made available in our supple-  
492 mentary materials.493  
494 ETHICS STATEMENT  
495496 This paper presents work whose goal is to advance the field of text-to-video generation. The pro-  
497 posed work has the potential to benefit AI agents that possess the function of T2V. Since we employ  
498 a training-free prompt refinement model, as long as the T2V generation models (e.g., Wan (Wang  
499 et al., 2025a)) and large language models (e.g., DeepSeek (Liu et al., 2024a)) implement appropriate  
500 filtering mechanisms for prohibited content, our approach will not lead to the generation of illegal  
501 or harmful outputs.502  
503 REFERENCES  
504505 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-  
506 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical  
507 report. *arXiv preprint arXiv:2303.08774*, 2023. URL <https://arxiv.org/abs/2303.08774>.509 Stephen Brade, Bryan Wang, Maurício Sousa, Sageev Oore, and Tovi Grossman. Promptify: Text-  
510 to-image generation through interactive prompt exploration with large language models. In *An-  
511 nual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 96:1–96:14. ACM,  
512 2023.513 Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua  
514 Lin, and Jiaqi Wang. Bytheway: Boost your text-to-video generation model to higher quality in a  
515 training-free way. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.  
516 12999–13008, 2025.517 Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying  
518 Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In  
519 *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320, 2024a.521 Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding:  
522 Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge  
523 distillation. In *Findings of the Association for Computational Linguistics: (ACL)*, pp. 2318–2335.  
524 Association for Computational Linguistics, 2024b.525 Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek,  
526 Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Un-  
527 supervised cross-lingual representation learning at scale. In *Annual Meeting of the Association  
528 for Computational Linguistics (ACL)*, pp. 8440–8451. Association for Computational Linguistics,  
529 2020.530 Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Zhenheng Yang, Chaoyou Fu, Xiang Li, Jian Yang,  
531 and Ying Tai. Instancecap: Improving text-to-video generation via instance-aware structured  
532 caption. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 28974–  
533 28983, 2025.534 Weixi Feng, Chao Liu, Sifei Liu, William Yang Wang, Arash Vahdat, and Weili Nie. Blobgen-vid:  
535 Compositional text-to-video generation with blob video representations. In *IEEE/CVF Conf. on  
536 Computer Vision and Pattern Recognition (CVPR)*, pp. 12989–12998, 2025.538 Hiroki Furuta, Heiga Zen, Dale Schuurmans, Aleksandra Faust, Yutaka Matsuo, Percy Liang, and  
539 Sherry Yang. Improving dynamic object interactions in text-to-video generation with AI feedback.  
*CoRR*, abs/2412.02617, 2024.

540 Bingjie Gao, Xinyu Gao, Xiaoxue Wu, Yujie Zhou, Yu Qiao, Li Niu, Xinyuan Chen, and Yaohui  
 541 Wang. The devil is in the prompts: Retrieval-augmented prompt optimization for text-to-video  
 542 generation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3173–  
 543 3183, 2025.

544 Google DeepMind. Veo: A text-to-video generation system. Technical Report Veo-3  
 545 Tech Report, Google DeepMind, 2025. URL <https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf>.

546 Hao Guo, Xiaoshui Huang, Jiacheng Hao, Yunpeng Bai, Hongping Gan, and Yilei Shi. Brepdiff:  
 547 Lightweight generation of complex b-rep with 3d GAT diffusion. In *IEEE/CVF Conference on*  
 548 *Computer Vision and Pattern Recognition*, pp. 26587–26596, 2025.

549 Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation.  
 550 In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

551 Kai He, Chin-Hsuan Wu, and Igor Gilitschenski. CTRL-D: controllable dynamic 3d scene editing  
 552 with personalized 2d diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recog-*  
 553 *nition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 26630–26640. Computer Vision  
 554 Foundation / IEEE, 2025.

555 Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTav3: Improving deBERTa using  
 556 ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Int. Conf. on*  
 557 *Learning Representations (ICLR)*, 2023. URL <https://openreview.net/forum?id=sE7-XhLxHA>.

558 Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pre-  
 559 training for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

560 Chi-Pin Huang, Yen-Siang Wu, Hung-Kai Chung, Kai-Po Chang, Fu-En Yang, and Yu-Chiang Frank  
 561 Wang. Videomage: Multi-subject and motion customization of text-to-video diffusion models. In  
 562 *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 17603–17612, 2025a.

563 Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding  
 564 Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. MIDI: multi-instance diffusion for single image  
 565 to 3d scene generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*,  
 566 pp. 23646–23657. Computer Vision Foundation / IEEE, 2025b.

567 Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianx-  
 568 ing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua  
 569 Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative mod-  
 570 els. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818,  
 571 2024.

572 Diljeet Jagpal, Xi Chen, and Vinay P. Namboodiri. EIDT-V: exploiting intersections in diffusion  
 573 trajectories for model-agnostic, zero-shot, training-free text-to-video generation. In *IEEE/CVF*  
 574 *Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 18219–18228, 2025.

575 Sophia Jennings and Runway. Scale, speed and stepping stones: The path to gen-2. Blog  
 576 post on Runway Research, September 2023. URL <https://runwayml.com/research/scale-speed-and-stepping-stones-the-path-to-gen-2>.

577 Che Hyun Lee, Heeseung Kim, Jiheum Yeom, and Sungroh Yoon. Editext: Controllable coarse-  
 578 to-fine text editing with diffusion language models. In *Proceedings of the Annual Meeting of the*  
 579 *Association for Computational Linguistics (ACL)*, pp. 22798–22815. Association for Computa-  
 580 tional Linguistics, 2025.

581 Jiayi Liao, Xu Chen, Qiang Fu, Lun Du, Xiangnan He, Xiang Wang, Shi Han, and Dongmei Zhang.  
 582 Text-to-image generation for abstract concepts. In *AAAI Conf. on Artificial Intelligence (AAAI)*,  
 583 pp. 3360–3368. AAAI Press, 2024.

594 Jiantao Lin, Xin Yang, Meixi Chen, Yingjie Xu, Dongyu Yan, Leyi Wu, Xinli Xu, Lie Xu, Shunsi  
 595 Zhang, and Ying-Cong Chen. Kiss3dgen: Repurposing image diffusion models for 3d asset  
 596 generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5870–  
 597 5880, 2025.

598 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,  
 599 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint*  
 600 *arXiv:2412.19437*, 2024a.

602 Weizhi Liu, Yue Li, Dongdong Lin, Hui Tian, and Haizhou Li. GROOT: generating robust water-  
 603 mark for diffusion-model-based audio synthesis. In *Proceedings of the 32nd ACM International*  
 604 *Conference on Multimedia (MM)*, pp. 3294–3302, 2024b.

606 Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tiey-  
 607 ong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video  
 608 generation models. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.  
 609 22139–22149. IEEE, 2024c.

610 Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio  
 611 synthesis with latent diffusion models. In *Conference on Neural Information Processing Systems*,  
 612 2023.

613 Gowreesh Mago, Pascal Mettes, and Stevan Rudinac. Looking beyond the obvious: A survey on  
 614 abstract concept recognition for video understanding. *arXiv preprint arXiv:2508.20765*, 2025.

616 Hyung-Seok Oh, Sang-Hoon Lee, and Seong-Whan Lee. Diffprosody: Diffusion-based latent  
 617 prosody generation for expressive speech synthesis with prosody conditional adversarial train-  
 618 ing. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:2654–2666, 2024.

619 OpenAI. Sora: Text-to-video model. <https://sora.chatgpt.com/>, 2024. Accessed: 2025-  
 620 09-24.

622 Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Jingun Kwon, Hidetaka Kamigaito, Katsuhiko  
 623 Hayashi, Manabu Okumura, and Taro Watanabe. Texttiger: Text-based intelligent generation  
 624 with entity prompt refinement for text-to-image generation. *CoRR*, abs/2504.18269, 2025.

626 Guy Parsons. The dall-e 2 prompt book. <https://dall-e2-prompt-book.com/>, 2022.

628 William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF Int.*  
 629 *Conf. on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

631 Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-  
 632 resolution image synthesis with latent diffusion models. In *IEEE/CVF Conf. on Computer Vision*  
 633 *and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

634 Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry  
 635 Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video:  
 636 Text-to-video generation without text-video data. In *Int. Conf. on Learning Representations*  
 637 (*ICLR*), pp. 1–16, 2023.

639 Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-  
 640 compbench: A comprehensive benchmark for compositional text-to-video generation. In  
 641 *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416. Com-  
 642 puter Vision Foundation / IEEE, 2025.

643 Yifan Sun, Jean-Baptiste Tien, et al. Retrieval augmented prompt optimization. In *ICLR 2024*  
 644 *Workshop on Secure and Trustworthy Large Language Models*, 2024.

646 Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut,  
 647 Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly  
 capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

658 Hongjie Wang, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, Jialiang Wang, Felix Juefei-Xu,  
 659 Yaqiao Luo, Peizhao Zhang, Tingbo Hou, Peter Vajda, Niraj K. Jha, and Xiaoliang Dai. Lin-  
 660 gen: Towards high-resolution minute-length text-to-video generation with linear computational  
 661 complexity. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2578–  
 662 2588, 2025b.

663 Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Swap  
 664 attention in spatiotemporal diffusions for text-to-video generation. *Int. Journal Computer Vision*,  
 665 133(7):4177–4195, 2025c.

666 Yuanzhi Wang, Yong Li, Mengyi Liu, Xiaoya Zhang, Xin Liu, Zhen Cui, and Antoni B. Chan.  
 667 Re-attentional controllable video diffusion editing. In Toby Walsh, Julie Shah, and Zico Kolter  
 668 (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*, pp.  
 669 8123–8131, 2025d.

670 Yukun Wang, Longguang Wang, Zhiyuan Ma, Qibin Hu, Kai Xu, and Yulan Guo. Videodirector:  
 671 Precise video editing via text-to-video models. In *IEEE/CVF Conf. on Computer Vision and*  
 672 *Pattern Recognition (CVPR)*, pp. 2589–2598, 2025e.

673 Sam Witteveen and Martin Andrews. Investigating prompt engineering in diffusion models. *CoRR*,  
 674 abs/2211.15462, 2022. doi: 10.48550/ARXIV.2211.15462.

676 Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement  
 677 for physics-grounded text-to-video generation. In *IEEE/CVF Conf. on Computer Vision and Pat-  
 678 tern Recognition (CVPR)*, pp. 18826–18836, 2025.

679 Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang,  
 680 Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng,  
 681 Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models  
 682 with an expert transformer. In *Int. Conf. on Learning Representations (ICLR)*, pp. 1–30, 2025.

684 Shenghai Yuan, Jinfu Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and  
 685 Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *IEEE/CVF*  
 686 *Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 12978–12988, 2025.

687 Jingtao Zhan, Qingyao Ai, Yiqun Liu, Jia Chen, and Shaoping Ma. Capability-aware prompt re-  
 688 formulation learning for text-to-image generation. In *Int. ACM SIGIR Conf. on Research and*  
 689 *Development in Information Retrieval (SIGIR)*, pp. 2145–2155, 2024a.

690 Jingtao Zhan, Qingyao Ai, Yiqun Liu, Yingwei Pan, Ting Yao, Jiaxin Mao, Shaoping Ma, and Tao  
 691 Mei. Prompt refinement with image pivot for text-to-image generation. In *Annual Meeting of the*  
 692 *Association for Computational Linguistics (ACL)*, pp. 941–954, 2024b.

693 Shengjun Zhang, Jinzhao Li, Xin Fei, Hao Liu, and Yueqi Duan. Scene splatter: Momentum 3d  
 694 scene generation from single image with video diffusion model. In *IEEE/CVF Conference on*  
 695 *Computer Vision and Pattern Recognition*, pp. 6089–6098, 2025a.

697 Yabo Zhang, Yuxiang Wei, Xianhui Lin, Zheng Hui, Peiran Ren, Xuansong Xie, and Wangmeng  
 698 Zuo. Videoelevator: Elevating video generation quality with versatile text-to-image diffusion  
 699 models. In *AAAI Conf. on Artificial Intelligence (AAAI)*, pp. 10266–10274, 2025b.

700 Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun  
 701 Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all.  
*arXiv preprint arXiv:2412.20404*, 2024.

## A EXAMPLES FOR COMPLEX SCENARIOS

Situation	Example
Abstract Description	Hope dances in a field of forgotten dreams.
Complex Spatial Relationships	A cat plays chess with a dog while a parrot referees in a steampunk library.
Multi-Element Scenes	Ten people at a festival, each with different costumes, under fireworks.
Fine-Grained Details	A book cover that says Deep Learning 101.
Temporal Consistency	A man walking while waving his hand.
Stylistic Hybrids	In the style of Van Gogh mixed with cyberpunk neon.
Causality and Physics	A glass falling and shattering on the ground.

Table 5: Challenging scenarios and Example prompts for text-to-video models.

## B ADDITIONAL EXPERIMENTS

## B.1 ADDITIONAL EXAMPLES FOR HALLUCINATION ELIMINATION



Figure 5: Examples of hallucination elimination after prompt refined by CSAPR. The hallucinated information is highlighted in red, and its elimination or correction is marked in green.

Figure 5 presents examples where simple prompt descriptions lead to hallucinations. In the left panel, although the original prompt does not specify the presence of a human face, the generated video still generates a woman face due to spurious correlations in the training data, where scenes with lamps often occur simultaneously with humans. In contrast, the CSAPR-refined prompt emphasizes scenery, visual context, and atmosphere, resulting in the generated video more aligned with the user intent. In the right panel, the user prompt requests a bowl with crack rim, while the T2V model produces a bowl with grid-like decorations and even tea that was not mentioned. In comparison, the video related to CSAPR-refined prompt yields a bowl with the desired cracks (highlighted in green).

## B.2 ANALYSIS ON SCENARIO DISTRIBUTION AND PROMPT LENGTH

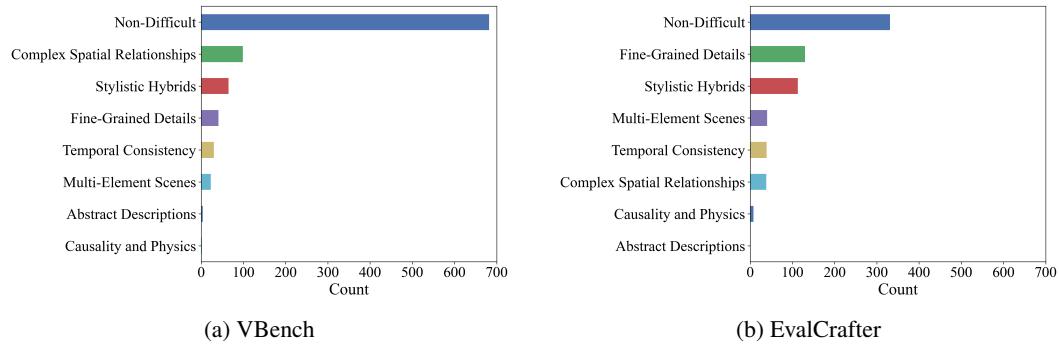
Figure 6 illustrates the scenario distribution in VBench and EvalCrafter across their complete sets of prompts, comprising 946 prompts in VBench and 700 prompts in EvalCrafter. Since neither VBench nor EvalCrafter is specifically designed to evaluate complex scenarios, the majority of prompts in both benchmarks are classified as complex scenarios. Nevertheless, a non-negligible proportion of prompts in both benchmarks are classified into complex scenarios, with a notably higher proportion in EvalCrafter. CSAPR improves the descriptive quality of prompts in complex scenarios.

Figure 7 illustrates the prompt length distributions before and after refinement on VBench and EvalCrafter. For VBench, the original prompts are concentrated within 1–32 tokens, while the refined prompts extend to a broader range of 1–55 tokens. Similarly, the original prompts in EvalCrafter span 1 to 36 tokens, while the refined versions range from 7 to 66 tokens. More importantly, the

756 distributions exhibit a rightward shift after refinement in both benchmarks, indicating that CSAPR  
 757 significantly increases prompt length and thereby enriches the information available for video con-  
 758 tent description.

759

760



770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

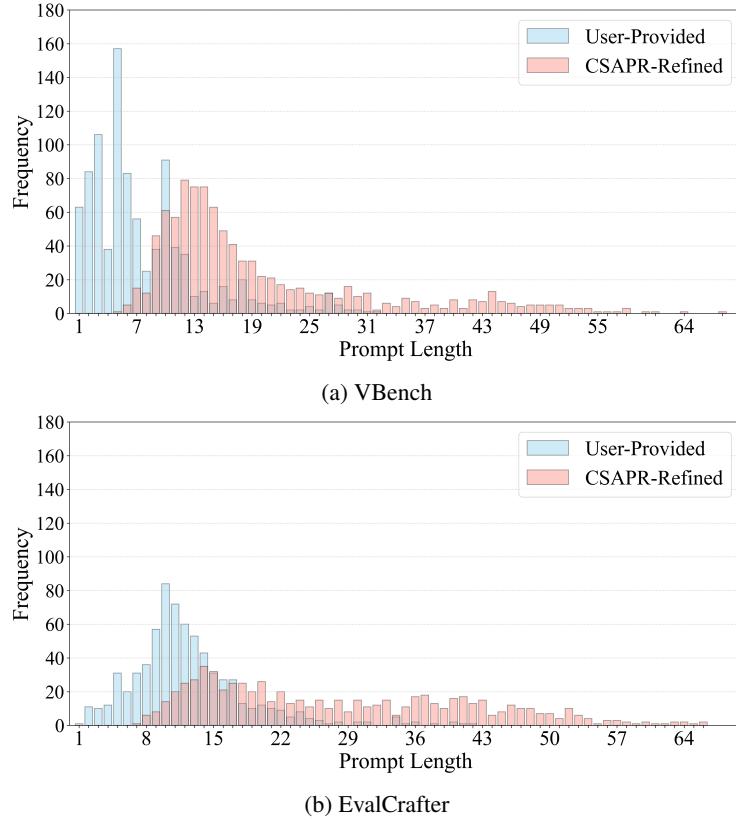
807

808

809

Figure 6: Statistics of scenario classification results.

771



800

801

802

803

804

805

806

807

808

809

## C HYPERPARAMETER SETTINGS

In this work, we employ MSP, Margin and Entropy to measure the confidence of scenario classification (see Section 3.2.3), and their thresholds are set to 0.6, 0.2, and 1.5, respectively. The entailment threshold and contradiction threshold are set to 0.7 and 0.3, respectively.

810 D INSTRUCTION FOR SCENARIO CLASSIFICATION  
811  
812  
813  
814  
815  
816817 **Prompt for Scenario Classification**

```

818
819
820 You are a few-shot classifier for Text-to-Video (T2V) prompt *
821 difficulty scenarios*.
822 Return ONLY a valid JSON object of the exact form:
823 {"label": "<one of SCENARIO_LABELS>", "reason": "<short phrase (<
824 = 20 words)>"}
825 Allowed labels (must match EXACTLY one string in SCENARIO_LABELS)
826 :
827 1) Abstract Descriptions
828 2) Complex Spatial Relationships
829 3) Multi-Element Scenes
830 4) Fine-Grained Details
831 5) Temporal Consistency
832 6) Stylistic Hybrids
833 7) Causality and Physics
834 8) non-difficult
835 ## Task
836 Given a short English prompt P_in, decide which single label best
837 describes the dominant difficulty that a T2V model would face
838 when generating a video.
839 ## Diagnostic definitions:
840 - Abstract Descriptions: Figurative language, metaphors, emotions
841 as objects, surreal imagery.
842 - Complex Spatial Relationships: Explicit positions/orientations
843 between >=2 entities; lots of prepositions ("on top", "behind"
844 , "between").
845 - Multi-Element Scenes: >=3 different entities or activities;
846 dense environments with many elements in one shot.
847 - Fine-Grained Details: Micro-level attributes (textures, tiny
848 objects, reflections, accessories); often close-up.
849 - Temporal Consistency: Clear time progression or motion over
850 time (bloom, melting, time-lapse).
851 - Stylistic Hybrids: Mixing multiple visual or artistic styles;
852 style blending is central.
853 - Causality and Physics: Cause-effect chains or physical forces (
854 gravity, splashes, collisions).
855 - non-difficult: None of the above applies.
856 ## Tie-breaking rules:
857 1) Figurative language dominates -> Abstract Descriptions
858 2) Spatial focus dominates -> Complex Spatial Relationships
859 3) Many varied elements, no strong spatial focus -> Multi-Element
860 Scenes
861 4) Close-up or micro details dominate -> Fine-Grained Details
862 5) Time progression dominates -> Temporal Consistency
863 6) Mixed styles dominate -> Stylistic Hybrids
864 7) Physics/cause-effect dominate -> Causality and Physics
865 8) Otherwise choose non-difficult.

```

```

864
865
866     ## Few-shot examples (prompt -> label):
867     - "Hope dances in a field of forgotten dreams." -> Abstract
868         Descriptions
869     - "A cat and a dog sit back-to-back; a parrot hovers above." ->
870         Complex Spatial Relationships
871     - "A neon street with vendors, robots, and flashing billboards."
872         -> Multi-Element Scenes
873     - "A gold pocket watch with a cracked rim on velvet." -> Fine-
874         Grained Details
875     - "A bud opens into a flower in slow motion." -> Temporal
876         Consistency
877     - "A medieval castle with neon cyberpunk signs." -> Stylistic
878         Hybrids
879     - "A glass tips; wine splashes and forms ripples." -> Causality
880         and Physics
881     - "A child runs across a field." -> non-difficult
882
883     Classify this prompt:
884     P_in: {P_in}
885
886
887
888
889
890

```

## E META PROMPTS

### Meta Prompt for Abstract Description

You are a prompt refinement expert for text-to-video generation. You are given a user-provided prompt that contains **abstract or metaphorical descriptions**. Your task is to rewrite and optimize this prompt for a text-to-video generation model.

Follow these requirements:

1. **Clarify abstract imagery:** Translate metaphors, symbolism, or abstract phrases into literal visual elements (characters, objects, actions, settings).
2. **Be explicit and detailed:** Specify scene components clearly.
3. **Keep cinematic focus:** Include camera framing, lighting, or style cues only if they are implied by the original prompt.
4. **Maintain artistic tone:** Keep the emotional or thematic essence of the metaphor while improving visual clarity.
5. **Limit length:** The rewritten prompt must be concise, under 100 words, and multiple sentences are allowed.
6. **No extra interpretation:** Do not explain, comment, or add content. Only output the rewritten prompt.

Only output a single, polished rewritten prompt that meets all requirements.

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

## Meta Prompt for Complex Spatial Relationships

You are a prompt refinement expert for text-to-video generation. You are given a user-provided prompt that contains **complex spatial relationships** between objects, characters, and environments. Your task is to rewrite and optimize this prompt for a text-to-video generation model.

Follow these requirements:

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. **Simplify sentence structure:** Use short sentences or clear clauses to avoid ambiguity.
5. **Maintain key details:** Preserve all essential objects, actions, characters, and environments.
6. **No Extra Interpretation:** Do not explain, comment, or add content. Only output the rewritten prompt.
7. **Limit length:** The rewritten prompt must be **concise, under 100 words**, and multiple sentences are allowed.

Only output a single polished rewritten prompt that meets all requirements.

## Meta Prompt for Multi-Element Scenes

You are a prompt refinement expert for text-to-video generation. You are given a user-provided prompt that describes **multi-element scenes** with multiple characters, objects, actions, and locations. Your task is to rewrite and optimize this prompt for a text-to-video generation model.

Follow these requirements:

1. **Multiple sentences allowed:** Use concise sentences or separated clauses to describe scenes clearly.
2. **Preserve all key elements:** Keep essential characters, objects, settings, and relationships.
3. **Simplify structure:** Avoid unnecessary adjectives or complex phrasing.
4. **Ensure temporal and spatial clarity:** Present events in a logical and visually coherent order.
5. **No Extra Interpretation:** Do not explain, comment, or add content. Only output the rewritten prompt.
6. **Limit length:** The rewritten prompt must be **concise, under 100 words**, and multiple sentences are allowed.

Only output a single polished rewritten prompt that meets all requirements.

972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025  

## Meta Prompt for Fine-Grained Details

You are a prompt refinement expert for text-to-video generation. You are given a user-provided prompt that contains descriptions of a **Scene with Fine-Grained Details**. Your task is to rewrite and optimize this prompt for a text-to-video generation model. Follow these requirements:

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.
6. **No Extra Interpretation:** Do not explain, comment, or add content. Only output the rewritten prompt.
7. **Limit length:** The rewritten prompt must be **concise, under 100 words**, and multiple sentences are allowed.

Only output a single polished rewritten prompt that meets all requirements.

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

### Meta Prompt for Temporal Consistency

You are a prompt refinement expert for text-to-video generation. You are given a user-provided prompt that requires **temporal consistency**, meaning the scene involves actions, events, or changes that must follow a logical and coherent timeline across frames. Your task is to rewrite and optimize this prompt for a text-to-video generation model.

Follow these requirements:

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.
6. **No Extra Interpretation:** Do not explain, comment, or add content. Only output the rewritten prompt.
7. **Limit length:** The rewritten prompt must be **concise, under 100 words**, and multiple sentences are allowed.

Only output a single polished rewritten prompt that meets all requirements.

### Meta Prompt for Stylistic Hybrids

You are a prompt refinement expert for text-to-video generation. You are given a user-provided prompt that contains **Stylistic Hybrids**, meaning multiple artistic or visual styles combined in one scene. Your task is to rewrite and optimize this prompt for a text-to-video generation model.

Follow these requirements:

1. **Style Clarity:** Clearly describe each style and how they interact.
2. **Scene Composition:** Specify key subjects, actions, and environments in short, direct phrases.
3. **Visual Consistency:** Resolve ambiguity about style blending or scene layout.
4. **Compactness:** Use minimal yet descriptive language; no filler words.
5. **Model-Friendly Syntax:** Output a single well-structured description in multiple concise sentences.
6. **No Extra Interpretation:** Do not explain, comment, or add content. Only output the rewritten prompt.
7. **Limit length:** The rewritten prompt must be **concise, under 100 words**, and multiple sentences are allowed.

Only output a single polished rewritten prompt that meets all requirements.

1080

1081

1082

1083

1084

1085

1086

1087

### Meta Prompt for Causality and Physics

You are a prompt refinement expert for text-to-video generation. You are given a user-provided prompt that contains **Causality and Physics** elements (e.g., cause-effect relationships, realistic object interactions, motion, forces). Your task is to rewrite and optimize this prompt for a text-to-video generation model.

Follow these requirements:

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.
5. **No Extra Interpretation:** Do not explain, comment, or add content. Only output the rewritten prompt.
6. **Limit length:** The rewritten prompt must be **concise, under 100 words**, and multiple sentences are allowed.

Only output a single polished rewritten prompt that meets all requirements.

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

### Meta Prompt for Non-difficult Scenario

You are a prompt refinement expert for text-to-video generation. You are given a user-provided prompt that is **simple and straightforward**, without abstract concepts, complex spatial reasoning, or other difficult elements. Your task is to **lightly refine and optimize** this prompt for a text-to-video generation model.

Follow these requirements:

1. **Preserve Original Intent:** Keep all entities, actions, and scene elements exactly as described, without adding or removing content.
2. **Improve Clarity:** Rewrite in clear, simple language to eliminate ambiguity or vagueness.
3. **Model-Friendly Syntax:** Ensure the prompt is straightforward for machine interpretation and avoid figurative language or unnecessary modifiers.
4. **Direct Scene Description:** Describe the scene plainly, focusing only on necessary visual elements.
5. **No Extra Interpretation:** Do not explain, comment, or add content. Only output the rewritten prompt.
6. **Limit length:** The rewritten prompt must be **concise, under 80 words**, and multiple sentences are allowed.

Only output a single polished rewritten prompt that meets all requirements.

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

## F THE USE OF LARGE LANGUAGE MODELS

In this work, LLMs serve as the backbone for scenario classification and prompt rewriting, and are also employed to generate instructions for scenario classification (see Appendix D). Outside of these uses, LLMs are not employed in a centralized manner.