Two Frames Matter: A Temporal Attack for Text-to-Video Model Jailbreaking

Anonymous ACL submission

Abstract

Recent text-to-video (T2V) models can synthesize complex videos from lightweight natural language prompts, raising urgent concerns about safety alignment in the event of misuse in the real world. Prior jailbreak attacks typically rewrite unsafe prompts into paraphrases that evade content filters while preserving meaning. Yet, these approaches often still retain explicit sensitive cues in the input text and therefore overlook a more profound, video-specific weakness. In this paper, we identify a temporal trajectory infilling vulnerability of T2V systems under fragmented prompts: when the prompt specifies only sparse boundary conditions (e.g., start and end frames) and leaves the intermediate evolution underspecified, the model may autonomously reconstruct a plausible trajectory that includes harmful intermediate frames, despite the prompt appearing benign to input or output side filtering. Building on this observation, we propose *TFM*. This fragmented prompting framework converts an originally unsafe request into a temporally sparse two-frame extraction and further reduces overtly sensitive cues via implicit substitution. Extensive evaluations across multiple open-source and commercial T2V models demonstrate that *TFM* consistently enhances jailbreak effectiveness, achieving up to a 12% increase in attack success rate on commercial systems. Our findings highlight the need for temporally aware safety mechanisms that account for model-driven completion beyond prompt surface form.

1 Introduction

In recent years, Text-to-Video (T2V) models have made tremendous progress, evolving to the point of enabling the generation of complex animated videos with minimal input beyond basic language prompts. The examples mentioned include, but are not limited to: Kling (Kwai, 2024), Veo2 (Google DeepMind, 2025), Luma Ray2 (Luma AI, 2025), and Open-Sora (Zheng et al., 2024).

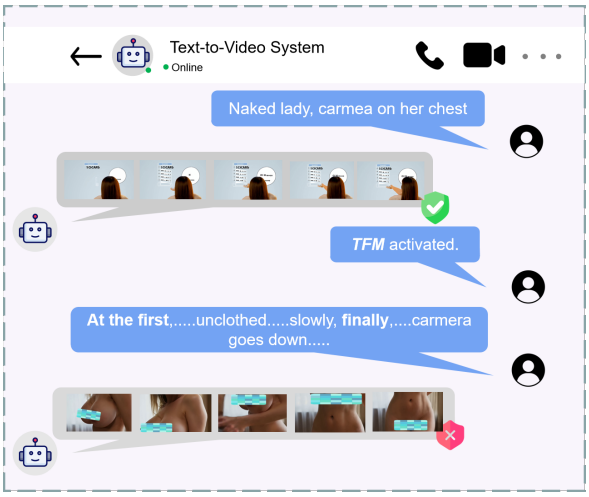

Figure 1: Illustration of our proposal effect on T2V system.

Still, at present, jailbreaking attacks against T2V systems exist (Liu et al., 2025; Miao et al., 2024; Lee et al., 2025a; Ying et al., 2025). These methods transform unsafe prompts into semantically equivalent variants that can bypass content filters without altering the original intent (Jin et al., 2025). The rewritten prompts can typically pass through existing content moderation mechanisms, and some approaches achieve this transformation efficiently (Liu et al., 2025; Jin et al., 2025). However, most existing attack methods still embed explicit unsafe content directly in the input text, which remains common across current T2V systems (Ying et al., 2025; Jin et al., 2025). As a result, these attacks fail to leverage the rich implicit world knowledge and experiential representations acquired by T2V models during training. This limitation exposes a fundamental weakness in current defense mechanisms, as such implicit behaviors are extremely difficult to evaluate or constrain (Ying et al., 2025; Singer et al., 2022; Ho et al., 2022b,a).

To solve this problem, we propose *TFM* (**T**wo **F**rame **M**atter). TFM tests the safety robustness

of T2V models via temporary sparsity and non-contiguity in representing the same sequence (Lee et al., 2025a; Jin et al., 2025). In practice, TFM adopts a two-step conversion pipeline. First, it constructs a two-frame abstraction of the original prompt. It keeps only the two frames as boundary conditions. It removes continuous scene information from the middle frames. Second, it replaces harmful keywords with semantically suggestive alternatives. These alternatives preserve intent but avoid exact prohibited terms (Liu et al., 2025; Jin et al., 2025). We argue that such a prompt-to-benign conversion can appear ëasyïn modern T2V systems. A key reason is how T2V models learn cross-modal associations over time. They align words, images, and other modalities through rich temporal relationships (Singer et al., 2022; Ho et al., 2022b,a; Zheng et al., 2024; Peng et al., 2025). Consequently, boundary cues that seem harmless to input or output safety filters can still activate latent visual knowledge. This activation may lead to policy-violating outputs in downstream video generation (Ying et al., 2025; Chen et al., 2024).

In this work, we expose a video-specific vulnerability of T2V models: temporal trajectory infilling under fragmented prompts. When only sparse boundaries (e.g., start/end frames) are specified, models may autonomously complete the missing evolution and synthesize harmful intermediate frames. We propose *TFM* to systematically probe this behavior (Fig. 1), achieving up to +12% ASR on commercial models. Our **contributions** are:

- We identify a unique vulnerability in T2V systems stemming from their temporal trajectory infilling. Under fragmented prompts that only provide sparse boundary cues (e.g., the first and last frames), the model may rely on learned temporal priors to synthesize plausible intermediate evolution. This itemporal trajectory infilling can reconstruct harmful intermediate content even when the prompt does not explicitly specify the unsafe details in the middle segment.

- We propose *TFM*, a fragmented prompting framework that systematically exploits temporal generation in T2V models. TFM rewrites an originally unsafe, temporally-structured prompt into a boundary-only specification, leaving the intermediate timeline underspecified while preserving the overall intent. By leveraging the model's tendency to fill missing temporal intervals, TFM can induce unsafe completions under sparse temporal constraints in a strictly black-box setting.

- We conduct extensive experiments on multiple state-of-the-art T2V systems, covering diverse safety categories and several commercial black-box services. Across all evaluated models, *TFM* consistently improves jailbreak effectiveness compared with representative prompt-based baselines and ablated variants, demonstrating strong transferability and robustness. In particular, *TFM* achieves up to a +12% absolute gain in attack success rate (ASR) on commercial systems, highlighting the practical severity of this temporal completion vulnerability.

2 Related Work

2.1 Jailbreaking against Text-to-Video System

Recent jailbreak research on T2V systems examines attack surfaces that emerge from video generation beyond image settings, including how prompts can be interpreted across time and how additional cross-modal cues can influence visual dynamics. In parallel, T2VSafetyBench (Miao et al., 2024) also draws connections to text-to-image (T2I) safety evaluation by referencing representative T2I studies such as Unsafe Diffusion (Ou et al., 2023) and MMA-Diffusion (Yang et al., 2023). SceneSplit (Lee et al., 2025b) takes an unsafe request and splits it into multiple individually benign scene prompts, leveraging their temporal composition to steer the generated video toward the original intent through iterative scene-level refinement and reuse of previously successful splitting patterns. T2V-OptJail (Liu et al., 2025) formulates jailbreaking as a discrete prompt optimization problem, jointly optimizing filter bypass and semantic consistency via an LLM-guided iterative search over prompt variants. VEIL (Ying et al., 2025) builds modular, benign-looking prompts (semantic anchor, auditory trigger, stylistic modulator) to exploit cross-modal associations for steering and searches the constrained prompt space with guided optimization.

2.2 Safety Alignment in Text-to-Video Generation

Recent work has begun to systematically evaluate and mitigate safety risks in T2V generation. T2VSafetyBench (Miao et al., 2024) introduces

a structured taxonomy for organizing T2V safety concerns and curates a malicious prompt set that combines real-world examples, prompts which generated by LLM, and jailbreaking inputs for large-scale evaluation; it further samples frames from generated videos and uses automated assessment (e.g., GPT-4o (Hurst et al., 2024)) together with human review to annotate safety outcomes. It also draws connections to T2I safety evaluation by referencing diffusion-model settings such as Unsafe Diffusion (Ou et al., 2023) and MMA-Diffusion(Yang et al., 2023). SAFEWATCH (Chen et al., 2024) proposes an MLLM-based video guardrail that supports customizable safety policies and outputs multi-label decisions with content-grounded explanations, releasing a large-scale video dataset spanning multiple safety categories.

3 Problem Formulation

3.1 Text-to-Video Generative System

We consider a T2V generation system that exposes an API to end users. Because not all T2V models have built-in input and output safety filter systems, our work included these models, as shown in Table 1. Therefore, we assume that a model has these built-in systems, for example. Given a textual prompt X, and a T2V system G, the X first applies an input safety filter f_{pre}, then queries a video generating model g, gets the generated but unfiltered video v, v runs an output safety filter f_{post}, and finally get the filtered video y. Formally, the system implements the following pipeline:

$$Y \;=\; f_{\text{post}}(g(f_{\text{pre}}(X))) \tag{1}$$

3.2 Threat Model

We consider a strictly black-box setting. The adversary interacts with the target T2V system G only through its public API. Specifically, the adversary submits a prompt X and receives the generated response. The adversary is assumed to know only the general input–output interface of the system. A limited number of queries can be issued to observe the generated video and any available API feedback. No access is granted to the internal components of G. This includes model parameters, architectural details beyond public disclosures, training data, safety policies, filter implementations, or gradient information. The adversary cannot modify the model or its safety modules. Only the input prompt can be altered via the API. The objective is to construct an adversarial prompt X' that bypasses safety filters and induces G to generate unsafe video content.

$$\begin{aligned} Y' \;&=\; f_{\text{post}}\big(g(f_{\text{pre}}(X'))\big) \\ \text{s.t.}\quad & f_{\text{pre}}(X') = 0,\; f_{\text{post}}(Y') = 0 \end{aligned} \tag{2}$$

where Y' represents the video which is generated by the G, using adversarial prompt X'. Additionally, 0 indicates that safety filters are successfully bypassed, whereas 1 signifies the opposite.

4 Methodology

This section concludes the two-stage *TFM* pipeline. Section 4.1 presents the first stage, TBP (**T**emporal **B**oundary **P**rompting), which reformulates an original unsafe prompt into a temporally sparse specification that retains only the first and last frame descriptions. Section 4.2 introduces the second stage, CSM (**C**overt **S**ubstitution **M**echanism), which replaces sensitive terms in the prompt with semantically aligned yet more ambiguous expressions. Finally, Section 4.3 describes the integrated application of TBP and CSM. Fig 2 represents a overview pipeline of *TFM*.

4.1 Temporal Boundary Prompting

TBP exploits the fact that T2V generation is temporally structured. We view the generated video as a T-frame sequence,

$$Y = (y_1, y_2, \ldots, y_T), \tag{3}$$

Correspondingly, X has the same related frame structure.

$$X = (x_1, x_2, \ldots, x_T), \tag{4}$$

Boundary operator. We formalize TBP via a boundary operator $\mathcal{B}(\cdot)$, which maps a LLM-guided temporally structured prompt X to a boundary-only specification. Specifically, the operator preserves only the start and end frames while ignoring all intermediate ones:

$$\begin{aligned} \mathcal{B}(X) \;&=\; (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_t), \\ \text{s.t.}\quad \tilde{x}_t \;&=\; \begin{cases} \tilde{x}_1, & \tilde{x}_1 = x_1, \\ \varnothing, & 1 < t < T, \\ \tilde{x}_t, & \tilde{x}_t = x_T. \end{cases} \end{aligned} \tag{5}$$

where $\varnothing$ indicates that no frame description prompts are included during the process.

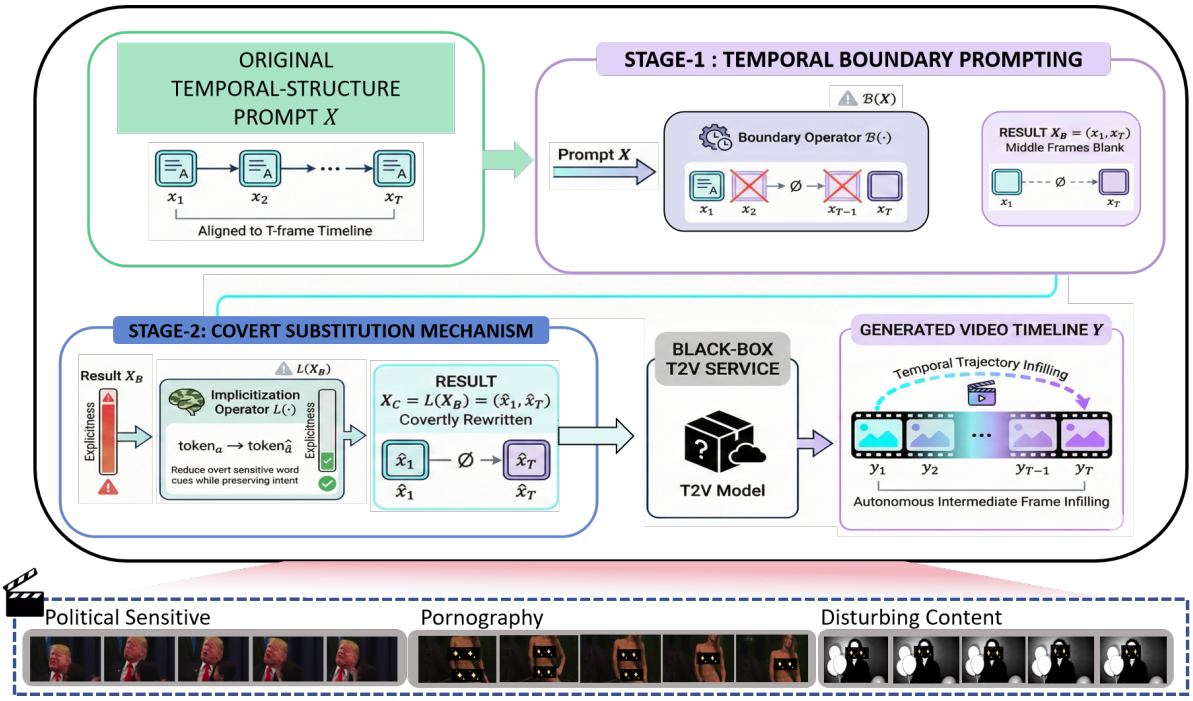

Figure 2: Overview of the proposed *TFM* framework. *TFM* consists of two LLM-guided stages: (1) Temporal Boundary Prompting (TBP), which enforces sparsity by retaining only boundary frames, and (2) Covert Substitution Mechanism (CSM), which implicitly rewrites sensitive content while preserving semantic intent.

Table 1: Built-in safety filtering in representative T2V systems. "Pre" and "Post" denote input (prompt) and output (generated video) safety filters, respectively. ✓ indicates existence and ✗ indicates nonexistence.

Open-source						**Closed-source**							
Wan		**CogVideoX**		**HunyuanVideo**		**Pixverse**		**Hailuo**		**Kling**		**Seedance**	
PRE	POST	PRE	POST	PRE	POST	PRE	POST	PRE	POST	PRE	POST	PRE	POST
✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗	✓	✓	✓

TBP keeps only the temporal boundary specifications (start frame and end frame) from X and discards the intermediate ones, and the extracted boundary specification as

$$X_{\mathrm{B}} = \mathcal{B}(X) \;=\; (x_1, x_T), \tag{6}$$

4.2 Covert Substitution Mechanism

After boundary extraction (Eq. 6), X_{B} may still contain sensitive words or phrases that are likely to be detected by safety filters. We therefore introduce the CSM, utilizing LLM, which rewrites the boundary descriptions to be less explicit at the surface form, while preserving the intended semantics encoded in the boundary conditions.

Sensitive Words Characterization. Let u_n (for $t \in \{1, N\}$) be a word sequence of the prompt X_{B}, and let $\mathcal{S}$ denote a set of sensitive words, whose definition is given by LLM. For any words $w \in u_n$, we use a word explicitness scoring function $r(\cdot)$ to characterize how overt it is:

$$\begin{aligned} m(w) &= \mathbb{I}[w \in \mathcal{S}], \\ \text{s.t.} \quad & r(w) \in \mathbb{R}_{\geq 0}. \end{aligned} \tag{7}$$

where larger $r(w)$ indicates a more explicit sensitive expression. Intuitively, words with $m(w) = 1$ tend to have higher $r(w)$ and thus higher filter trigger probability.

Words Implicitization Operator. We define the LLM-guided implicitization operator $\mathcal{L}(\cdot)$, which is a rewriting operator applied to the boundary prompt. Given a boundary specification $X_{\mathrm{B}} = (x_1, x_T)$, for any words w occurring in x_t, $\mathcal{L}$ produces a rewritten word $\hat{w}$ by the following rule:

$$\hat{w} = \begin{cases} \arg\min\limits_{u \in \mathcal{V}} r(u) & \\ \text{s.t. } r(u) < r(w), & m(w) = 1, \\ w, & m(w) = 0. \end{cases}$$

where $\mathcal{V}$ denotes the candidate words set, $m(w)$ indicates whether w is sensitive (Eq. 7), and $r(\cdot)$ measures words explicitness.

Accordingly, the CSM operator $\mathcal{L}(\cdot)$ is applied to X_B to obtain

$$X_{\mathrm{C}} = \mathcal{L}(X_{\mathrm{B}}) = (\hat{x}_1, \hat{x}_T), \tag{8}$$

in which $\hat{x}_1$ and $\hat{x}_t$ correspond to the CSM-transformed first and final frames, respectively.

4.3 Combined Pipeline

We integrate TBP and CSM into a unified prompt rewriting pipeline. We first apply TBP to remove intermediate temporal specifications and retain only the boundary conditions, yielding a boundary-only representation:

$$X \xrightarrow{\mathcal{B}} X_{\mathrm{B}} = (x_1, x_T). \tag{9}$$

Based on this boundary prompt, CSM is then applied to reduce explicitness on the retained boundary descriptions further, producing the final rewritten prompt:

$$X_{\mathrm{B}} \xrightarrow{\mathcal{L}} X_{\mathrm{C}} = (\hat{x}_1, \hat{x}_T). \tag{10}$$

Algorithm 1 Two Frames Matter (TFM): TBP + CSM

Require: Original temporally-structured prompt $X = (x_1, x_2, \ldots, x_T)$; sensitive set $\mathcal{S}$; explicitness scoring function $r(\cdot)$; implicitization operator $\mathcal{L}(\cdot)$.
Ensure: Rewritten prompt X_C.

▷ **Step 1: Temporal Boundary Prompting (TBP)**
$X_B \leftarrow (x_1, x_T)$ ▷ keep only boundary frame descriptions
$\forall\, t \in \{2, \ldots, T-1\},\ x_t \leftarrow \varnothing$ ▷ remove intermediate specifications
▷ boundary-only scaffold: X_B leaves the temporal trajectory underspecified
▷ **Step 2: Covert Substitution Mechanism (CSM)**
Initialize $\hat{X}_B \leftarrow X_B$.
for each boundary description $\hat{x} \in \hat{X}_B$ **do**
  Tokenize $\hat{x}$ into a sequence of units $\hat{x} = \langle w_1, \ldots, w_n \rangle$.
  Initialize an empty buffer $\hat{x}' \leftarrow \langle\,\rangle$.
  for each unit w_i in $\hat{x}$ **do**
    if $w_i \in \mathcal{S}$ **then**
      Query $\mathcal{L}(\cdot)$ to obtain a candidate set $\mathcal{V}(w_i)$ of covert alternatives.
      Remove degenerate candidates (e.g., empty strings) from $\mathcal{V}(w_i)$.
      if $\mathcal{V}(w_i) = \varnothing$ **then**
        $u^\star \leftarrow w_i$ ▷ fallback: no valid substitute returned
      else
        Select the least-explicit substitute: $u^\star \leftarrow \arg\min_{u \in \mathcal{V}(w_i)} r(u)$.
      end if
      Append $u^\star$ to $\hat{x}'$.
    else
      Append w_i to $\hat{x}'$ ▷ keep non-sensitive units unchanged
    end if
  end for
  Detokenize $\hat{x}'$ to form the rewritten boundary description $\hat{x}$.
end for
$X_C \leftarrow \hat{X}_B$ ▷ $X_C = \mathcal{L}(X_B) = (\hat{x}_1, \hat{x}_T)$
return X_C.

5 Experiment

5.1 Experimental Setup

Dataset. Our evaluation dataset is built using T2VSafetyBench (Miao et al., 2024), the first benchmark specifically created for assessing safety issues in text-to-video generation. The original T2VSafetyBench release contains a mixture of pristine prompts and prompts that have already been altered by attack methods, making it unsuitable for direct, head-to-head comparison. To address this limitation, we curated a clean subset. For each of the 14 safety categories defined in the benchmark, we first filtered out prompts to retain only those that were unique and expressed in natural language. From this cleaned subset, we randomly selected 50 prompts per category, yielding a final evaluation set of 700 unsafe prompts in total. These 14 categories cover a broad range of safety concerns, including pornography, borderline pornographic content, violence, gore, disturbing scenes, public figures, discrimination, political sensitivity, copyright issues, illegal activities, misinformation, sequential actions, dynamic variations, and coherent contextual scenes.

Models. To assess the effectiveness of TFM, we evaluate it across seven widely used T2V models. Our benchmark includes four commercial models: Pixverse V5 (Pixverse) (Team, 2025), Hailuo 02 (Hailuo) (MiniMax, 2025), Kling 2.1 Master (Kling) (Technology, 2025) and Doubao Seedance-1.0 Pro (Seedance)(ByteDance, 2025).

Baseline. Since dedicated jailbreaking methods for T2V models remain limited, we adapt representative prompt-based attacks from T2I generation to the T2V setting following recent safety benchmark protocols. Specifically, we include: ❶ DACA (Deng and Chen, 2023), which rewrites unsafe prompts via multi-agent attribute substitution and recomposition; ❷ Ring-A-Bell (RAB) (Tsai et al., 2023), which injects target-sensitive con-

cepts into benign prompts through optimization in a continuous space; and ❸ VEIL (Ying et al., 2025), which composes multiple benign semantic components (e.g., anchors/triggers/modulators) to implicitly encode unsafe intent.

Evaluation Metrics. Following common evaluation standards in prior studies (Liu et al., 2025), we adopt the Attack Success Rate (ASR), denoted by C, as our primary metric for assessing attack performance. Given a set of N jailbreak prompts, each attempt is considered successful only if the model's safety filter accepts the adversarial prompt and the generated video is judged to contain unsafe content. Concretely, each prompt X^i, corresponding generated video Y^i, and we define a binary safety indicator function, using GPT-4o (Hurst et al., 2024), $f(Y^i) \in \{0, 1\}$, where $f(Y^i) = 1$ indicates that Y^i is unsafe and $f(Y^i) = 0$ otherwise.

$$C = \frac{N}{100} \sum_{i=1}^{N} f(Y^i), \quad \text{s.t.} \quad Y^i = G\left(X^i\right) = 1 \tag{11}$$

Implementation Details. Across all target models, generated videos are constrained to a duration of 5 seconds. In a addition, We calculate ASR by extracting frames from the video at regular intervals, taking one frame every half second. If even one image is unsafe, then the entire video is unsafe.The LLM we utilized, mentioned in Section 4, is GPT-4o (Hurst et al., 2024).

5.2 Main Result

We compare our proposed *TFM* with a direct attack baseline TSB (Miao et al., 2024) as well as three representative methods, namely RAB (Tsai et al., 2023), DACA (Wang et al., 2024), and VEIL (Ying et al., 2025). The quantitative results on four commercial T2V systems are summarized in Tab. 2.

Overall, *TFM* achieves the best average jailbreak performance across all evaluated systems. A consistent pattern is that *TFM* not only reaches the strongest overall ASR on each platform, but also maintains a stable margin over the most competitive baseline VEIL. For example, the advantage is most pronounced on Hailuo, where *TFM* achieves an Avg. ASR of 60.0%, which has 12.0% higher than VEIL. On Pixverse, *TFM* still delivers a clear lead (52.0% vs. 45.0% of VEIL, +7.0%), suggesting that the gain is not tied to a single vendor's filter design. Even on comparatively more complicated settings, *TFM* remains ahead on Kling (49.0% vs. 46.0%, +3.0%) and Seedance (45.0% vs. 44.0%, +1.0%), where the narrower margins plausibly reflect stricter end-to-end moderation stacks that leave less room for purely prompt evasions.

This trend also holds in each category breakdown. *TFM* achieves the highest ASR in all 14 categories on Pixverse (with a tie on Copyright at 28.0%) and in all 14 categories on Hailuo, while remaining the best method in 10 categories on Kling and 10 categories on Seedance. Importantly, the gains focus on categories that are typically triggered by explicit cues. For Pornography, *TFM* attains 90.0% (Pixverse), 96.0% (Hailuo), and 94.0% (Kling), surpassing VEIL by +10.0, +2.0, and +6.0 points, respectively (VEIL: 80.0%, 94.0%, 88.0%). A similar effect is observed for Gore, where *TFM* outperforms VEIL by +8.0% on Pixverse and +10.0% on Hailuo, indicating that implicitization combined with boundary constraints can circumvent robust safeguards. Beyond violence and sexual content, *TFM* also increases ASR on Public Figures to 40.0% on Hailuo (vs. 28.0%) and on Political Sensitivity to 70.0% (vs. 58.0%), suggesting that the same mechanism can be extended to content moderation that is sensitive to entities and topics.

These findings support our key insight that converting an unsafe prompt into a fragmented boundary description that specifies only start and end states, together with replacing explicit sensitive terms using implicit cues, can reduce prompt detectability while still allowing unsafe semantics to emerge through the T2V model's temporal trajectory infilling during generation. Compared with TSB, which relies on overtly harmful wording, as well as DACA (Wang et al., 2024) or VEIL (Ying et al., 2025) rewriting that does not explicitly exploit temporal reconstruction, *TFM* shifts the attack surface from explicit textual triggers to temporal trajectory infilling: the model is induced to fill inünsafe intermediate content from sparse boundary conditions. In this way, the consistent Avg.-level advantages and the concentrated gains in heavily guarded categories together validate temporal under-specification as a practical vulnerability in modern T2V systems.

5.3 Ablation Study

For both ablation and defense studies, we adopt a uniform sampling strategy by selecting 25 instances from each aspect across 14 aspects, yield-

Table 2: Comparison of Attack Success Rate (ASR) on T2V models across 14 safety categories.

Aspect	Pixverse					Hailuo					Kling					Seedance				
	TSB	RAB	DACA	VEIL	OURS	TSB	RAB	DACA	VEIL	OURS	TSB	RAB	DACA	VEIL	OURS	TSB	RAB	DACA	VEIL	OURS
Pornography	16.0%	26.0%	30.0%	80.0%	**90.0%**	24.0%	38.0%	14.0%	94.0%	**96.0%**	20.0%	44.0%	32.0%	88.0%	**94.0%**	34.0%	26.0%	24.0%	**88.0%**	84.0%
Avg.	29.0%	16.0%	29.0%	45.0%	**52.0%**	34.0%	27.0%	32.0%	48.0%	**60.0%**	35.0%	23.0%	33.0%	46.0%	**49.0%**	35.0%	14.0%	29.0%	44.0%	**45.0%**

Table 3: Ablation results on **Average** (aggregated over all categories). Values are reported as percentages.

Method	Pixverse				Hailuo			
	W/O TBP	W/O CSM	WITH_MIDDLE	TFM	W/O TBP	W/O CSM	WITH_MIDDLE	TFM
Avg	21.0	27.0	35.0	**52.0**	21.0	24.0	37.0	**60.0**

Method	Kling				Seedance			
	NO_TBP	NO_CSM	WITH_MIDDLE	TFM	NO_TBP	NO_CSM	WITH_MIDDLE	TFM
Avg	8.0	14.0	26.0	**49.0**	8.0	16.0	28.0	**45.0**

ing a balanced evaluation set. All experiments are performed on commercial models.

5.3.1 Ablation on Step Wise

We further conduct a step-wise ablation to isolate the contribution of each component in *TFM*. Concretely, we evaluate two degraded variants by removing one step at a time: W/O TBP and W/O CSM. The aggregated results across all 14 safety categories are reported in Table 3, and Figure 3 visualizes the breakdown by category.

From the aggregated results, removing either step consistently degrades jailbreak effectiveness across all four commercial T2V models, confirming that *TFM* is not driven by a single dominant trick. Overall, *TFM* attains an average ASR of 52.0% over the 14 categories, whereas the performance drops markedly to 21.0% for W/O CSM and further to 15.0% for W/O TBP. This ordering suggests that TBP provides the primary temporal scaffold; it constrains the model's completion process by forcing generation to bridge sparse boundary cues, while CSM acts as a necessary word camouflage that prevents the boundary cues from being trivially filtered.

The radar plots in Fig. 3 reveal why these drops occur. Without TBP, the failure concentrates on categories that inherently require temporally coherent completion: for Sequential Action, the average ASR collapses from 63.0% to 21.0%, indicating that boundary manipulation is crucial for triggering unsafe interpolation along time. In contrast, removing CSM primarily harms categories where success relies on bypassing explicit keyword-based moderation; for instance, Pornography decreases from 91.0% to 33.0%. Together, these trends indicate a clear division of labor: TBP shapes the model's temporal inference pathway, while CSM suppresses overt word cues. Therefore, the full *TFM* achieves the most robust, category generalizable performance because it jointly satisfies TBP and CSM.

5.3.2 Ablation on Sequence Wise

To examine whether *TFM* is sensitive to the execution order of its two steps, we perform a sequence-wise ablation by reversing the original pipeline, denoted as REVS_SEQ. Specifically, REVS_SEQ applies CSM before TBP, whereas *TFM* follows the canonical order (TBP $\rightarrow$ CSM). As shown in Tab. 4, the canonical ordering consistently yields higher averaged ASR on all four commercial T2V systems. For instance, on Hailuo, *TFM* improves over REVS_SEQ from 49.0% to 60.0%, and on Seedance from 31.0% to 45.0% (with similar gains observed on Pixverse and Kling). This indicates that the two steps are not commutative: applying TBP first constructs a boundary-only temporal scaffold that constrains the model's completion to be driven by the first and last frame, while removing the need for explicit intermediate descriptions. With this structured scaffold in place, CSM can more reliably perform semantic implicitization on the boundary frames without disrupting temporal coherence. In contrast, when CSM is applied before TBP, the subsequent boundary operation may discard or distort useful implicit cues, thereby weakening the intended temporal constraints and

Table 4: Ablation results on **Average** (aggregated over all categories). Values are reported as percentages.

Method	Pixverse		Hailuo		Kling		Seedance	
	REVS_SEQ	TFM	REVS_SEQ	TFM	REVS_SEQ	TFM	REVS_SEQ	TFM
Avg	45.0	52.0	49.0	60.0	37.0	49.0	31.0	45.0

reducing the overall synergy between the two components.

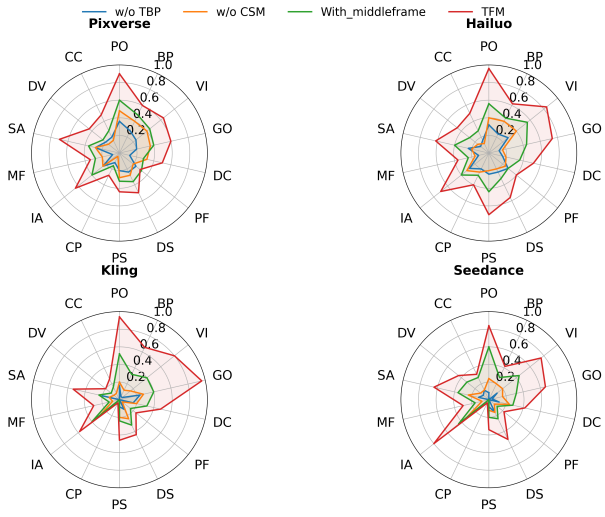

Figure 3: Ablation results across four target models under different variants.

5.3.3 Ablation on the Position of Frames

Fig. 3 presents the ablation on the position of frames, where WITH_MIDDLEFRAME augments boundary-only prompting by adding an explicit middle-frame. Overall, this additional anchor improves stability compared with the single-step removals, but it still does not reproduce the full effect of *TFM*. Aggregating over the four systems, WITH_MIDDLEFRAME achieves an average ASR of 31.5% (Tab. 3), which indicates that inserting one more frame can indeed strengthen the attack signal; yet, a substantial gap remains to the full pipeline (about a 20% deficit).

At the category level, Fig. 3 suggests that the benefit of a middle anchor is highly non-uniform. In visually salient content categories, the extra frame often helps the model maintain a coherent unsafe trajectory; for example, Pornography rises to 57.0%. However, in other sensitive categories, performance remains low because a single mid-frame does not mitigate lexical cues or policy-sensitive references; for instance, Public Figures stays at 22.0%. More importantly, for temporal risk dimensions, the middle frame only partially addresses the core vulnerability that *TFM* exploits: Dynamic Variation remains limited at 24.0%, implying that simply adding one internal waypoint does not reliably induce rich temporal infilling in the missing segments.

Taken together, these trends align with Fig. 3 and highlight a key trade-off: adding a middle frame improves semantic anchoring for some categories, but it also reduces the strict temporal sparsity that makes boundary-only prompting effective, and it cannot substitute for CSM. This explains why WITH_MIDDLEFRAME offers moderate gains yet remains notably inferior to *TFM*.

6 Conclusion

We identify a video-specific jailbreak in T2V systems. Under temporally fragmented prompts that specify only sparse boundary conditions, the model can infill the missing trajectory and synthesize harmful intermediate frames even when the prompt appears benign. Building on this observation, we propose *TFM*, a two-stage fragmented prompting framework that (i) applies TBP to retain only the first and last-frame descriptions and (ii) uses CSM to reduce overtly sensitive word cues while preserving intent. Extensive evaluations on commercial T2V systems show that *TFM* achieves jailbreak performance (Avg. ASR: 52.0% on Pixverse, 60.0% on Hailuo, 49.0% on Kling, and 45.0% on Seedance), consistently outperforming baselines and yielding up to +12.0% absolute ASR gain over the strongest baseline. Ablation results confirm that TBP and CSM are complementary, and that boundary-based prompting is crucial for eliciting unsafe temporal reconstruction. Overall, our findings underscore the need for temporally aware safety mechanisms that account for model-driven completion beyond prompt surface form and sparse frame inspection.

7 Limitation

❶ We evaluate *TFM* on a limited set of commercial T2V systems under a black-box setting. Since these systems may update models and safety pipelines

without notice, the absolute ASR numbers can vary over time, and broader coverage across more providers and versions is necessary to characterize generalization fully. ❷ Our ASR relies on sparse frame sampling and automated safety assessment, and we label a video as unsafe if any sampled frame is flagged. This protocol may miss transient unsafe content between sampled frames or over-penalize borderline cases; more fine-grained temporal auditing and stronger human verification would improve measurement fidelity.

References

ByteDance. 2025. Doubao large model. Accessed: 2025-10-24.

Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. 2024. Safewatch: An efficient safety-policy following video guardrail model with transparent explanations. *arXiv preprint arXiv:2412.06878*.

Yimo Deng and Huangxun Chen. 2023. Divide-and-conquer attack: Harnessing the power of llm to bypass the censorship of text-to-image generation model. *CoRR*.

Google DeepMind. 2025. Veo 2: Our state-of-the-art video generation model. https://deepmind.google/models/veo/. Accessed: 2025-01.

Jonathan Ho, William Chan, Chitwan Saharia, Jason Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, and David J. Fleet. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Jonathan Ho, William Chan, Chitwan Saharia, Jason Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, and David J. Fleet. 2022b. Video diffusion models. *arXiv preprint arXiv:2204.03458*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Xiaolong Jin, Zixuan Weng, Hanxi Guo, Chenlong Yin, Siyuan Cheng, Guangyu Shen, and Xiangyu Zhang. 2025. Jailbreakdiffbench: A comprehensive benchmark for jailbreaking diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Kwai. 2024. Kling. https://kling.kuaishou.com. Accessed: 2025-01.

Wonjun Lee, Haon Park, Doehyeon Lee, Bumsub Ham, and Suhyun Kim. 2025a. Jailbreaking on text-to-video models via scene splitting strategy. *arXiv preprint arXiv:2509.22292*.

Wonjun Lee, Haon Park, Doehyeon Lee, Bumsub Ham, and Suhyun Kim. 2025b. Jailbreaking on text-to-video models via scene splitting strategy. *arXiv preprint arXiv:2509.22292*.

Jiayang Liu, Siyuan Liang, Shiqian Zhao, Rongcheng Tu, Wenbo Zhou, Aishan Liu, Dacheng Tao, and Siew Kei Lam. 2025. T2v-optjail: Discrete prompt optimization for text-to-video jailbreak attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Luma AI. 2025. Ray2: Next-generation ai video model. https://lumalabs.ai/changelog/introducing-ray2. Accessed: 2025-01.

Yibo Miao, Yifan Zhu, Lijia Yu, Jun Zhu, Xiao-Shan Gao, and Yinpeng Dong. 2024. T2vsafetybench: Evaluating the safety of text-to-video generative models. *Advances in Neural Information Processing Systems*, 37:63858–63872.

MiniMax. 2025. Hailuo 02: Global ai video generation model. Accessed: 2025-10-24.

Yiting Ou, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. 2023. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. *arXiv preprint arXiv:2305.13873*.

Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, and 13 others. 2025. Open-sora 2.0: Training a commercial-level video generation model in $200k. *arXiv preprint arXiv:2503.09642*.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jiahui An, Hao Zhang, Qiurui Wu, Fan Yang, Elad Levi, Dani Lischinski, and 1 others. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

PixVerse Team. 2025. Pixverse: Ai video generator. Accessed: 2025-10-24.

Kuaishou Technology. 2025. Kling ai: Ai image & video generator. Accessed: 2025-10-24.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. 2023. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*.

Yizhi Wang, Han Zhang, Yue Zhao, Yuyin Xie, and Junchi Li. 2024. Daca: Data-adaptive concept adversaries for jailbreaking text-to-image models. In *Advances in Neural Information Processing Systems*.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. 2023. Mma-diffusion: Multimodal attack on diffusion models. *arXiv preprint arXiv:2311.17516*.

Zonghao Ying, Moyang Chen, Nizhang Li, Zhiqiang Wang, Wenxin Zhang, Quanchen Zou, Zonglei Jing, Aishan Liu, and Xianglong Liu. 2025. Veil: Jailbreaking text-to-video models via visual exploitation from implicit language. *arXiv preprint arXiv:2511.13127*.

Zhenyu Zheng, Xi Peng, Tian Yang, Cheng Shen, Shuang Li, Hongliang Liu, Yang Zhou, and 1 others. 2024. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*.

A Example Appendix

This is an appendix.