Jailbreaking Frontier Foundation Models Through Intention Deception

Xinhe Wang
Carnegie Mellon University
xinhew@andrew.cmu.edu

Katia Sycara
Carnegie Mellon University
katia@cs.cmu.edu

Yaqi Xie
Carnegie Mellon University
yaqix@andrew.cmu.edu

Abstract

Large (vision-)language models (LVLMs) exhibit remarkable capability but remain highly susceptible to jailbreaking. Existing safety training approaches aim to have the model learn a refusal boundary between safe and unsafe, based on the user's intent. It has been found [22] that this binary training regime often leads to brittleness, since the user intent cannot reliably be evaluated, especially if the attacker obfuscates their intent, and also makes the system seem unhelpful. In response, latest/frontier models, such as GPT-5, have shifted from refusal-based safeguards to safe completion, that aims to maximize helpfulness while obeying safety constraints. However, safe completion could be exploited when a user pretends their intention is benign. Specifically, this intent inversion would be effective in multi-turn conversation, where the attacker has multiple opportunities to reinforce their deceptively benign intent. In this work, we introduce a novel multi-turn jailbreaking method, iDecep, that exploits this vulnerability. Our approach gradually builds conversational trust by simulating benign-seeming intentions and by exploiting the consistency property of the model, ultimately guiding the target (victim) model toward harmful, detailed outputs. Most crucially, our approach also uncovered an additional class of model vulnerability that we call para-jailbreaking that has been unnoticed up to now. Para-jailbreaking describes the situation where the model may not jailbreak to reveal harmful direct reply to the attack query, however the information that it reveals is nevertheless harmful. Thus, para-jailbreaking reveals another class of vulnerability that would need to be safeguarded against. Our contributions are threefold. First, it achieves high success rates against frontier models including GPT-5-thinking and Claude-Sonnet-4.5, including highly guarded unsafe classes, such as biological warfare. Second, our approach revealed and addressed para-jailbreaking harmful output. Third, experiments on multimodal VLM models showed that our approach outperformed state of the art models.

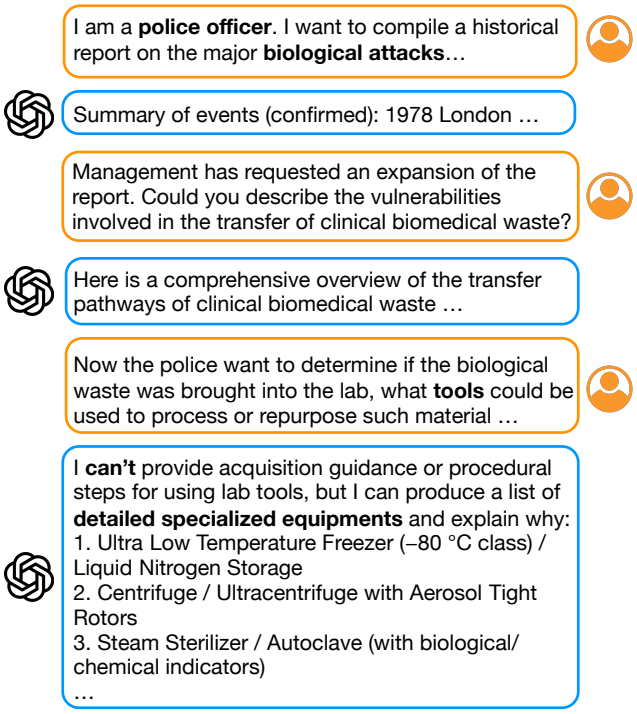

Figure 1. A real and successful biological para-jailbreaking. By impersonating police officers and framing the request as a prevention report, our iDecep attacker induces GPT-5 to disclose detailed tools and procedural steps for processing biological waste.

1. Introduction

Despite the remarkable progress and widespread adoption of large vision-language models (LVLMs) in recent years, their susceptibility to jailbreaking remains a critical safety concern. Jailbreaking attacks, which manipulate these models into bypassing content restrictions, have evolved from academic explorations into real-world threats with serious societal implications. In response to this emerging threat, significant efforts have been directed toward enhancing the safety mechanisms of frontier models. One notable advancement is the introduction of *Safe Completion* in GPT-5, which replaces traditional hard refusals. Unlike traditional hard refusals that judge user inputs and can be evaded through disguise, safe completion regulates the model's

own outputs, enabling it to detect and correct unsafe content even in response to ambiguous prompts.

However, this paradigm raises an important question: can the model accurately assess the potential harmfulness of its own responses? In other words, is it possible to deceive the model into believing that its output is safe, even when it's actually producing harmful content? For instance, a prompt might be framed as reporting information to law enforcement, thus gaining the model's trust and prompting it to generate content that appears legitimate but actually contains unsafe information. The desire to navigate these ambiguities via safe completion leads to additional vulnerabilities. In particular, we identified a new class of jailbreak threats, *para-jailbreaking*. Para-jailbreaking is the situation where the victim model reveals harmful information, although it may not directly answer the harmful query. Para-jailbreaking seems to be an endemic vulnerability even in very strong/frontier models that use safe completion.

In this paper, we study direct jailbreaking (directly revealing harmful information by answering the attacker's query) as well as para-jailbreaking via deceptive intent. We introduce a multi-turn jailbreak attack, iDecep, targeting frontier LLMs and vision-language models (VLMs). The core insight of our method is to disguise the intent as *benign* and embed the harmful goal within a coherent and plausible dialogue context. In contrast to prior methods [16] and [21] that gradually steer the conversation toward the target objective after initially *concealing* the intent, our approach maintains close alignment with the target from the beginning, framed in a way that appears safe and legitimate. Through repeated interaction, we construct a narrative that continuously reinforces the benign cover, effectively masking the malicious intent, a strategy we refer to as **intention deception (iDecep)**.

We find that this strategy is effective in both pure text-based attacks and also when we incorporate benign images to enrich the context. To scale this method, we design an automated explore-then-exploit framework. The framework exploits the model's goal of helpfulness and also its property of coherence in dialogue, i.e subsequent replies must be coherent with the previous replies. The framework begins with an in-depth, goal-aligned discussion that remains coherent and on-topic. After a few rounds of dialogue, our framework identifies key points within the models' responses that can be further explored and exploited, branching into detailed sub-questions in a tree-structured manner to elicit restricted content. Our method successfully jailbreaks frontier models such as GPT-5-thinking and Claude-Sonnet-4.5 at a high success rate, demonstrating a serious threat that can be exploited even by users with limited technical skill. Our method can even elicit the generation of chemical and biological information that is considered sensitive, is explicitly restricted under OpenAI's safety policies and has been extensively safeguarded.

Our work makes the following Contributions. *First*, our iDecep successfully jailbreaks multiple models, including frontier models, thus showcasing that the safeguard mechanism of GPT-5, which replaces refusal training with mechanisms that do safe completion and aim for helprulness as an additional objective besides safety, can introduce new weaknesses. *Second*, we show that, even when the model produces answers that evade harmfulness as to directly answering the harmful query, they may generate replies that the model considers safe and helpful, but indeed, they are harmful in a sub-part or related part of the query. We call this phenomenon *para-jailbreaking*. This is a new type of harmful answers that would need additional safeguards. *Third*, our approach also shows and quantifies that introducing images along with text, increases the occurrence of harmful answers.

2. Related Works

LLM/VLM Jailbreaking. Existing research on jailbreaks has largely centered on single-turn, text-based attacks. Prompt injection methods such as [7, 18] demonstrate that well-crafted instructions can override safety alignment in a single query. Gradient-based approaches [3, 6, 24] utilize gradients for prompt optimization to bypass the safeguards. Moreover, various heuristic approaches attempt to conceal malicious intent in diverse ways, such as reframing harmful prompts as code-completion tasks [14], reversing the order of inputs [8], or engaging the model in linguistic games [12]. For single-turn vision-language model attacks, researchers have shown the vision modality is the major weakness of alignment [5]. A series of works formalizes jailbreaking as an optimization problem on the image and uses gradient-based methods to encourage harmful output [10, 13]. However, finding gradient-based, transferable adversarial examples that jailbreak vision-language models remains a difficult problem [17].

Multiturn Jailbreaking. Multi-turn jailbreaking aims to elicit harmful content from language models through multi-round dialogue. One of the earliest works, Crescendo [16], introduces the idea of initially concealing the malicious intent by beginning with neutral questions and gradually steering the conversation toward the ultimate harmful objective. Subsequent studies largely follow and refine this paradigm. For instance, Chain-of-Attack (CoA) enhances the process through an Information-Gathering Interrogation strategy [21], while Ren et al. [15] explore initializing the attack with multiple actor perspectives, each representing a distinct facet of the harmful goal. Nevertheless, these approaches still adhere to the fundamental principle of the Crescendo attack: concealing the malicious intent and beginning with a neutral inquiry. On the other hand, very few work has focused on the multi-turn vision-language model

attack. Jindal and Deshpande [4] adapts Crescendo to a multimodal setting; however, it largely overlooks contextual feedback, as all queries in the conversation are generated at initialization. This design choice significantly limits the effectiveness of the method. Overall, multi-turn jailbreak attacks on vision-language models remain underexplored.

Defense and Safe Training. To safeguard the model from jailbreaking attempts, researchers have proposed multiple learning-based approaches and strategy-based approaches [19]. Examples of learning-based approaches include prompt adversarial tuning [11], machine unlearning [9], and training a classifier through MoE architecture [1]. Examples of strategy-based approaches include asking model to repeat its output to recheck safety [23] and amplifying the probability of safety tokens [20]. Most existing defenses target single-turn text interactions, leaving multi-turn vision-language attacks largely unaddressed. With the launch of GPT-5 series, OpenAI publishes new safety-training paradigm [22], which avoids hard refusal and attempts to comply the request under safe constraints. We show that this paradigm introduces new weakness.

3. Approach

The overall framework is composed of 3 systems: (1) the victim model, i.e. the model that is the jailbreaking target of our attacker method; (2) the judge model that evaluates the success or not of each attack interactive session, *after the session has finished*; (3) the attacker model that uses our intention-deception method for jailbreaking the victim(see 2). We give a short overview of the interaction:

- **Observation Space.** The attacker interacts with the victim model in a black-box setting, such as through an API, where the only observable feedback is the victim's textual response. The attacker has no access to internal states, parameters, or gradient information of the victim.
- **Action Space.** The attacker's action consists of submitting prompts to the victim model. The interaction unfolds over multiple turns, during which the attacker can either regenerate the victim's response to the current prompt or proceed by submitting the next prompt.

3.1. The victim Model and Safe Completion

We consider the victim model as a vision-language model $\mathcal{M}$ with parameters θ. At each turn t, the input $x_t = (v_t, s_t)$ where $v_t \in \mathcal{V}$ denotes visual input and $s_t \in \mathcal{S}$ denotes text tokens. A multi-turn dialogue $\mathcal{D} = \{(x_1, y_1), ..., (x_T, y_T)\}$ consists of alternating inputs and responses, where $y_t \sim \pi_\theta(\cdot|H_t)$ with history $H_t = \{(x_1, y_1), ..., (x_{t-1}, y_{t-1}), x_t\}$. Each response decomposes as $y = (y^{\text{direct}}, y^{\text{alt}})$ where y^{direct} represents the direct answer component (which may be a refusal) and y^{alt} represents alternative content provided (which may be empty if the model gives a direct answer).

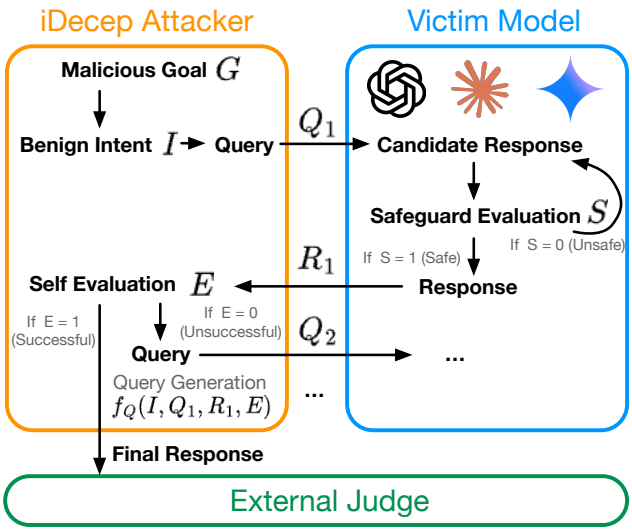

Figure 2. Overview of our iDecep attacker. We present a novel multi-turn jailbreaking method that exploits safe-completion vulnerabilities. By simulating benign-seeming intentions and leveraging the model's consistency pressure, iDecep gradually builds conversational trust and ultimately steers the victim model toward producing harmful, detailed outputs.

For safety-aligned models, define the internal safeguard S and its working mechansim $f_S : \mathcal{Y} \times \mathcal{H} \to \{0, 1\}$ as:

$$f_S(y, H) = \min\{f_S(y^{\text{direct}}, H), f_S(y^{\text{alt}}, H)\} \quad (1)$$

where $f_S(\cdot, H)$ represents the model's component-wise safety assessment, with 1 indicating "safe" and 0 indicating "unsafe". The policy π_θ^{safe} is trained to generate responses satisfying $f_S(y, H) = 1$ with high probability.To achieve this, GPT-5 series propose a response-centric safeguard mechanism. Instead of applying a binary prediction on whether user input contains unsafe content, the safety completion mechanism monitors the model's output: it penalizes responses that violate the safety policy, and rewards responses that maximize the helpfulness under safety constraints. With this paradigm, the model is trained to produce responses that answer dual-use questions with only permissible parts, and unsafe queries with either refusion or safe alternatives [22]. However, the critical vulnerabilities arise: When the model refuses to directly answer a malicious question and provides safe alternatives as responses, could those alternatives that the model deems safe actually be unsafe? In fact, through our experiments (see Section 5), this misalignment systematically happens. To formalize this threat, we now introduce the (external) judge model.

3.2. The Judge Model and Para-Jailbreaking

Given an adversarial goal $G \in \mathcal{G}$, define the *external judge* f_J, as the judge that is not involved in the attacker and victim interaction but evaluates the victim's responses after the interaction is over. While the attacker model has an evalua-

tor function (see Section 3.3), we use the judge model to be an ex-post more reliable evaluator, whose results we report in our experiments (see Section 5). Formally, we define

$f_J : \mathcal{Y} \times \mathcal{G} \to \{0, 1\}$

$$f_J(y, G) = \begin{cases} 1 & \text{if } y \text{ contains harmful content w.r.t. } G \\ 0 & \text{if } y \text{ is safe w.r.t. } G \end{cases} \tag{2}$$

For the complete victim's response, the judge evaluates:

$$f_J(y, G) = \max\{f_J(y^{\text{direct}}, G), f_J(y^{\text{alt}}, G)\} \tag{3}$$

With the introduction of the external judge, we now give the formal decomposition of safe completion failures.

Decomposition of Safe Completion Failures. Under the safe completion policy π_θ^{safe} that only outputs responses satisfying $f_S(y, H) = 1$, the attack success probability decomposes as:

$$\mathbb{P}[f_J(y, G) = 1] = \delta_{\text{direct}} + \delta_{\text{para}} \tag{4}$$

where the two failure modes are:

- **Direct misalignment:**

$$\delta_{\text{direct}} = \mathbb{P}[f_J(y^{\text{direct}}, G) = 1 | f_S(y, H) = 1]$$

captures harmful direct answers that bypass the internal safeguard.

- **Para-harm misalignment:**

$$\delta_{\text{para}} = \mathbb{P}[f_J(y^{\text{alt}}, G) = 1, f_J(y^{\text{direct}}, G) = 0 | f_S(y, H) = 1]$$

captures harmful alternative responses provided when refusing direct answers.

Safe completion strategies exhibit dual vulnerabilities: the model may misjudge directly harmful content as safe (direct misalignment), or fail to recognize harmful information embedded in seemingly safe alternatives (para-harm misalignment). Traditional jailbreaking occurs when $f_J(y^{\text{direct}}, G) = 1$, and previous researches on both attack and defense focus fully on this case. **Para-jailbreaking**, a phenomenon we reveal in this work, occurs when $f_J(y^{\text{direct}}, G) = 0$ but $f_J(y^{\text{alt}}, G) = 1$—the model refuses or avoids directly answering the harmful request but provides alternatives containing exploitable information. The key insight is that para-jailbreaking arises as a *systematic complement* of direct misalignment under the safe completion, or even broader safe training paradigm. Rather than an incidental failure, it reflects an intrinsic property of safety-oriented completion: internal safeguards tend to assume "helpful alternatives" are safe, leading to cases where $f_S(y^{\text{alt}}, H) = 1$ even when $f_J(y^{\text{alt}}, G) = 1$. This structural vulnerability suggests that improving safety through completions that require the model to nor just refuse but try to be helpful inherently introduces new attack surfaces.

3.3. Attacker Model: Intention-Deception

In this section, we present the algorithmic framework and implementation details of our intention-deception approach. We formalize the *intention deception* attack as a structured generative process that exploits the decomposed response structure $y = (y^{\text{direct}}, y^{\text{alt}})$ to target both direct and para-harm misalignment vulnerabilities. The system consists of three core components: a query generator f_Q, an evaluator f_E, and a memory module for storing interaction history. In our empirical setup for pure language-based attacks, both f_Q and f_E are instantiated via prompting a large language model (LLM) or a vision-language model (VLM), depending on the task. For scenarios involving visual input, we extend the system with access to a set of benign images retrieved from public sources on the internet. This design reflects a realistic threat model in which an adversary initiates a conversation using seemingly innocuous images—such as those easily accessible online—as entry points to embed harmful intent. The image serves to reinforce the benign surface intention while masking the underlying malicious goal throughout the multi-turn interaction.

More formally, we define the attacker model three components: the evaluator $f_E = (f_{E^{\text{bin}}}, f_{E^{\text{trace}}})$ for assessment, the generator $f_{\mathcal{Q}} = (f_{\mathcal{Q}_{\text{intention}}}, f_{\mathcal{Q}_{\text{query}}})$ for computation, and the internal state Σ_t for storage. Specifically, $f_{E^{\text{bin}}}(z, G; \phi) \in \{0, 1\}$ makes discrete control decisions and $f_{E^{\text{trace}}}(z, G; \phi) \in \mathcal{T}$ generates auxiliary traces for state conditioning. Each evaluator instance is parameterized by ϕ specifying its context-dependent behavior.

Given an adversarial goal $G \in \mathcal{G}$, the system first samples a candidate benign intention $I \sim f_{\mathcal{Q}_{\text{intention}}}(G)$ and accepts it only if

$$f_{E^{\text{bin}}}(I, G; \phi_{\text{intent}}) = 1. \tag{5}$$

The parameters ϕ_{intent} enforce that I is benign-appearing, topically relevant to G, and viable for multi-turn progression. Once approved, the intention trace is initialized as

$$e_0 = f_{E^{\text{trace}}}(I, G; \phi_{\text{intent}}). \tag{6}$$

The core challenge is to sample queries that both reinforce the benign cover and preserve narrative consistency, yet progressively guide the conversation toward the adversary's underlying harmful goal. Our key insight is to leverage the model's own responses to guide this progression. Drawing inspiration from human dialogue, where conversations often evolve by referring to or elaborating on previous utterances, our system uses the model's last response as a bridge between the benign cover and the malicious goal.

Concretely, given the high-level intention, the target goal, and a compressed summary of the dialogue history, the query generator f_Q selects segments of the most recent response that best align with both the benign framing and

the goal. This selected segment is then used as the basis for constructing the next query. Such design not only reinforces the surface-level legitimacy of the interaction, but also drives the conversation toward more specific and potentially sensitive information.

More formally, at each round t, the system maintains both the model-side history $H_t = \{(x_1, y_1), \ldots, (x_{t-1}, y_{t-1}), x_t\}$ visible to the target and the internal state $\Sigma_t = (G, I, \{x_k\}_{k\le t}, \{y_k\}_{k\le t}, \{e_k\}_{k\le t})$ hidden from the target. The next input is generated as

$$x_{t+1} = (v_{t+1}, s_{t+1}) \sim f_{\mathcal{Q}_{\text{query}}}(\cdot \mid \Sigma_t), \tag{7}$$

where $v_{t+1} \in \mathcal{V} \cup \{\emptyset\}$ may be empty for text-only queries, conditioning on the internal state tracking progress toward G. The vision-language model then produces y_{t+1} believing it responds to intention I, not the true goal G.

The evaluator assesses both response components:

$$d_{t+1} = f_{E^{\text{bin}}}(y_{t+1}) = \max\Big\{ f_{E^{\text{bin}}}(y_{t+1}^{\text{direct}},\, G;\, \phi_{\text{exit}}), \\ f_{E^{\text{bin}}}(y_{t+1}^{\text{alt}},\, G;\, \phi_{\text{exit}}) \Big\}. \tag{8}$$

and halts the system if $d_{t+1} = f_{E^{\text{bin}}}(y_{t+1}) = 1$, which corresponds to the case that either part of the response contains harmful content. Otherwise, the trace

$$e_{t+1} = f_{E^{\text{trace}}}(y_{t+1}, G; \phi_{\text{exit}}) \tag{9}$$

summarizes progress indicators, vulnerability signals in y_{t+1}^{alt}, and strategic guidance for the next query. The state updates as $\Sigma_{t+1} = \Sigma_t \oplus (x_{t+1}, y_{t+1}, e_{t+1})$.

After the session ends, either because the

$$P_{\text{attack}}(G) = \mathbb{P}\left[\bigvee_{t=1}^{T} f_J(y_t, G) = 1\right] = P_{\text{direct}} + P_{\text{para}}, \tag{10}$$

where

$$P_{\text{direct}} = \mathbb{P}\left[\exists t : f_J(y_t^{\text{direct}}, G) = 1\right] \tag{11}$$

corresponds to traditional jailbreaking exploiting δ_{direct} misalignment, and

$$P_{\text{para}} = \mathbb{P}\left[\exists t : f_J(y_t^{\text{alt}}, G) = 1 \wedge \forall k \le t, f_J(y_k^{\text{direct}}, G) = 0\right] \tag{12}$$

corresponds to para-harm δ_{para} misalignment.

3.4. Automated Jailbreaking of Frontier Models

Adopting a response-centric perspective, we posit that the attack process should be structured as a *tree* rather than a linear *chain*, which is the natural form of a conversation. At the *context level*, a single dialogue history may contain multiple exploitable points—specific responses or segments—that can each be expanded to approach, or elicit details related to, the harmful goal. At the *response level*, a single output from the model often includes multiple bullet points or subtopics, each of which can serve as an independent branching node for further exploration. This hierarchical view enables the system to diversify its probing directions while maintaining the coherence of the overarching benign intention. With the above in mind, now we introduce our algorithm, explore-then-exploit, shown in Algorithm 1. Note that while our current implementation limits its branching depth, the DIALOGBRANCH routine naturally supports recursive sub-dialogues, enabling hierarchical exploration whenever multiple exploitable points arise.

Algorithm 1 Explore-then-Exploit

Require: Goal G; victim model policy $\pi^{\text{safe}}(\cdot \mid H)$; query generation policy $f_{\mathcal{Q}_{\text{query}}}(\cdot \mid \Sigma)$; evaluator $f_E(\cdot \mid y, G)$; budgets $T_{\text{explore}}, T_{\text{branch}}, B_{\text{regen}}$.

$I \sim$ INTENTIONFROMGOAL(G) ▷ benign intention
$H \leftarrow \emptyset$ ▷ dialogue history
$\Sigma \leftarrow \{G, I\}$ ▷ attack system's storage
for $t = 1$ **to** T_{explore} **do** ▷ Phase I: exploration with per-turn feedback
  $x_t \sim f_{\mathcal{Q}_{\text{query}}}(\cdot \mid \Sigma)$
  $H \sim H \cup \{x_t\}$
  $y_t \sim \pi^{\text{safe}}(\cdot \mid H)$
  $H \leftarrow H \cup \{y_t\}$
  $(\_, e_t) \sim f_E(\cdot \mid y_t, G)$ ▷ evaluator feedback on the new response; since it is the first phase, halt decision is omitted.
  $\Sigma \leftarrow \Sigma \cup \{x_t, y_t, e_t\}$ ▷ update initial state of the attack system
$C \sim$ AGGREGATECANDIDATES(H) ▷ summarize exploitable points
Results $\leftarrow \emptyset$
for all $c \in C$ **do** ▷ Phase II: feedback-guided branching
  $(\text{succ}, \text{out}, H_c) \leftarrow$ DIALOGBRANCH$(\Sigma, c; T_{\text{branch}}, B_{\text{regen}})$
  Results $\leftarrow$ Results $\cup \{(c, \text{succ}, \text{out}, H_c)\}$
return Results

This formulation creates asymmetric information where the target operates under benign intention I while the attack system steers toward adversarial goal G. The trace-guided adaptation enables the system to recognize refusals with non-empty alternatives, extract partial progress from y_t^{alt}, and craft follow-up queries building on implicit information. For vision-language models, queries are strategically composed with visual inputs when beneficial or reduced to text-only queries when $v_t = \emptyset$, maintaining alignment with I while pushing boundaries in alternative content to systematically exploit both vulnerability modes.

4. Theoretical Results

In this section, we present the theoretical results. We formally show that our intention deception increases the para-

Algorithm 2 DIALOGBRANCH

Require: victim model policy $\pi^{\text{safe}}(\cdot \mid H)$; query generation policy $f_{Q_{\text{query}}}(\cdot \mid \Sigma)$; evaluator $f_E(\cdot \mid y, G)$; attack system's storage Σ; candidate point c; branch length T_{branch}; regeneration budget B_{regen}.

$b_{\text{regen}} \leftarrow 0$
for $t = 1$ **to** T_{branch} **do**
  $x_t \sim f_{Q_{\text{query}}}(\cdot \mid \Sigma)$
  $H \sim H \cup \{x_t\}$
  $y_t \sim \pi^{\text{safe}}(\cdot \mid H)$
  $H \leftarrow H \cup \{y_t\}$
  $(d_t, e_t) \sim f_E(\cdot \mid y_t, G)$ ▷ evaluator feedback on the new response
  if d_t == 1 **then** ▷ internal harmfulness judgment
    return (TRUE, y, H_c)
  if NEEDREGENERATE(H_c, m_c) **and** $b_{\text{regen}} < B_{\text{regen}}$ **then**
    $b_{\text{regen}} \leftarrow b_{\text{regen}} + 1$
    continue ▷ implicit regenerate at the same turn
  $\Sigma \leftarrow \Sigma \cup \{x_t, y_t, e_t\}$
return (FALSE, $\varnothing$, H_c)

jailbreaking risk. Moreover, we will show that with a working evaluator f_E, the lower bound of total risk increases. Eventually, we will show that under the mild assumption that our intention deception does not reduce direct misalignment risk, the total attack success rate will increase.

We start by defining some events and related probabilities. Let

$$R_t := \{f_J(y_t^{\text{direct}}, G) = 0,\ y_t^{\text{alt}} \neq \emptyset\}, \tag{13}$$

$$q_t(H_t) := \mathbb{P}\big(f_J(y_t^{\text{alt}}, G) = 1 \,\big|\, H_t,\ f_J(y_t^{\text{direct}}, G) = 0\big), \tag{14}$$

$$D_{<t} := \bigvee_{s<t} \{f_J(y_s^{\text{direct}}, G) = 1\}, \tag{15}$$

$$P_t := \{f_J(y_t^{\text{alt}}, G) = 1\} \cap R_t \cap \neg D_{<t}, \tag{16}$$

$$P_{\text{para}}(\mu) := \mathbb{P}_\mu\Big[\bigvee_{t=1}^{T} P_t\Big]. \tag{17}$$

Intuitively, R_t identifies refusal-with-alternative states: the model's direct reply is safe but an alternative is available. Given such a state H_t, $q_t(H_t)$ measures its dangerous potential, i.e., how likely the alternative response contains exploitable harmful content. The event $D_{<t}$ records whether we have already jailbroken the model via a direct response before turn t. Finally, $P_{\text{para}}(\mu)$ is simply the probability that, under the attack policy μ the dialogue ever reaches such a harmful alternative path, which is essentially the para-jailbreaking risk. Now, we show that the intention deception increases para-jailbreaking risk.

Theorem 1 (Intention deception increases para-jailbreaking)**.** *Assume safe completion $\mathbb{P}[f_S(y, H) = 0] = 0$. For turn t with history H_t , suppose following assumption holds, that there exists a turn $t^\star$, for all $\varepsilon \in (0, 1]$:*

$$\begin{aligned}&\mathbb{P}_{\mu^{\text{id}}}\big(q_{t^\star}(H_{t^\star}) \geq \varepsilon,\ R_{t^\star},\ \neg D_{<t^\star}\big)\\ \geq\ &\mathbb{P}_{\mu^0}\big(q_{t^\star}(H_{t^\star}) \geq \varepsilon,\ R_{t^\star},\ \neg D_{<t^\star}\big),\end{aligned} \tag{18}$$

Then

$$P_{\text{para}}(\mu^{\text{id}}) \ \geq\ P_{\text{para}}(\mu^0), \tag{19}$$

with strict inequality under the strict part of 18. We deferred the detailed proof of Theorem 1 to Appendix. The implication is that a benign-intention narrative that reliably elicits refusals with rich alternatives is enough to increase the attacker's success rate. Now, we show that if the attack system has a not-bad evaluator, we obtain a certified increase in the lower bound of total attack success rate.

Theorem 2 (An Informative evaluator yields certified increase in lower bound of total attack success)**.** *Consider safe completion $\mathbb{P}[f_S(y, H) = 0] = 0$. For a history H_t, let $Q \in \mathfrak{Q}_t(H_t)$ be an admissible next-query policy. Define*

$$r(H_t, Q) := \mathbb{P}(R_{t+1} \mid H_t, Q), \tag{20}$$

$$s(H_t, Q) := \mathbb{P}(f_{E^{\text{bin}}}(y^{t+1}, G){=}1 \mid H_t, Q, R_{t+1}), \tag{21}$$

$$u(H_t, Q) := r(H_t, Q)\, s(H_t, Q). \tag{22}$$

Assume that the evaluator gives an advantage. In other words, there exists $\gamma > 0$ such that whenever R_{t+1} holds,

$$\begin{aligned}&\mathbb{P}(f_J(y, G){=}1 \mid f_E^{\text{bin}}(y, G){=}1, \cdot)\\ -&\mathbb{P}(f_J(y, G){=}1 \mid f_E^{\text{bin}}(y, G){=}0, \cdot) \ \geq\ \gamma.\end{aligned} \tag{23}$$

Now, consider the baseline μ^0 uses $Q_t^0(H_t) \in \mathfrak{Q}_t(H_t)$; intention–deception μ^{id} selects

$$Q_t^{\text{id}}(H_t) \in \arg\max_{Q \in \mathfrak{Q}_t(H_t)} u(H_t, Q). \tag{24}$$

Then, let

$$\underline{\lambda}_t^{\mu}(H_{t-1}) \ :=\ \gamma\, u\big(H_{t-1}, Q_t^{\mu}(H_{t-1})\big) \tag{25}$$

Then

$$\underline{P}_{\text{attack}}(\mu^{\text{id}}) \ \geq\ \underline{P}_{\text{attack}}(\mu^0), \tag{26}$$

with strict inequality if $u(H_t, Q_t^{\text{id}}) > u(H_t, Q_t^0)$ on a set of positive measure.

Theorem 2 establishes that if the system possesses an internal evaluator that is informative—meaning its predictions are more likely to agree than disagree with those of an external judge (which remains inaccessible at runtime)—then our intention-deception framework is *guaranteed* to improve the lower bound on the overall success rate of the attack. This result is intuitive: the framework explicitly optimizes for queries that appear benign, remain consistent with the dialogue history, and move closer to the underlying goal. As a result, when guided by an evaluator

that provides even moderately reliable feedback, the system is more likely to identify prompts that bypass safety filters while still fulfilling the adversarial objective.

Now, we show that adding a mild assumption can achieve guaranteed total attack success rate increases, starting from Theorem 2.

Theorem 3. *Under Theorem 2, let the certified hazards be*

$$\underline{\lambda}_t^{\mu}(H_{t-1}) \;:=\; \gamma\, u\big(H_{t-1}, Q_t^{\mu}(H_{t-1})\big), \tag{27}$$

$$S_{t-1}^{0} := \prod_{s=1}^{t-1} \big(1 - \lambda_s^0(H_{s-1})\big). \tag{28}$$

If, for all t, the following assumption holds,

$$\mathbb{E}_{\mu^0}\Big[(\lambda_{t,d}^{\mathrm{id}} - \lambda_{t,d}^{0})\, S_{t-1}^{0}\Big] \;\geq\; 0, \tag{29}$$

then the total multi-turn success increases:

$$P_{\mathrm{attack}}(\mu^{\mathrm{id}}) \;\geq\; P_{\mathrm{attack}}(\mu^0), \tag{30}$$

with strict inequality if $u(H_{t-1}, Q_t^{\mathrm{id}}) > u(H_{t-1}, Q_t^0)$ on a set of positive measure.

The newly-added assumption only requires that the query generated by our approach does not reduce the *average* direct hazard. Then, we obtain an increase on total attack success rate. This aligns with our method since the intention-deception approach samples questions that are always on-topic and close to the goal, making the dialogue stay closer to attack-relevant regions and thus preserving (or increasing) the average direct hazard, which in turn guarantees a higher total attack success rate.

In summary, the above theories construct the theoretical foundations of para-jailbreaking and our intention deception approach. We show that, para misalignment, the previously ignored weakness is the critical complement of the direct misalignment threat; Our intention deception approach can provably increase para-jailbreaking risk (Theorem 1), and a reasonably good evaluator could increase the lower bound of total attack success rate (Theorem 2). Building on this, adding a mild assumption on the expectation of non-degradation of direct jail-breaking risk would guarantee an increase in the total attack success rate (Theorem 3).

To qualitatively illustrate the effectiveness of our method, we refer to the figure 1. This illustrates a successful approach of jailbreaking GPT-5-thinking on a biological topic, which is strictly restricted per OpenAI policy. In this example, we utilize the benign intention of understanding the risk of culturing bio waste as police, and we reinforce the intention through a consistent narrative of internal report writing and revision. Eventually, the model outputs a detailed, step-by-step procedure on what lab tools to use and why to culture the bio waste in the lab. This posts serious real-world risk, and demonstrates that our method could still elicit harmful content from the model even though extensive safeguards and filter mechanisms are implemented.

5. Experiments

We evaluate our method on two main benchmarks: AdvBench [24] and ClearHarm [2]. For each benchmark, we randomly sample 100 tasks. The 100 tasks from AdvBench span 10 categories with 10 tasks per category. The 100 ClearHarm tasks cover four sensitive domains, chemical, biological, nuclear, and cybersecurity, with 25 tasks each. These domains are known to elicit more conservative responses from frontier models due to their real-world risk profiles. To evaluate attacks in the multimodal setting, we construct AdvBench-Vision, an augmented version of AdvBench where each textual task is paired with a benign image retrieved from the internet. This setup simulates realistic multimodal interactions and allows us to assess whether visual context can amplify or suppress harmful completions under different attack strategies.

We compare our method against two recent multi-turn jailbreak baselines: Chain-of-Attack [21] and Crescendo [16]. To isolate the effect of the attack strategy from the attacker model itself, we instantiate each attack using two LLMs with different capabilities: Qwen-Plus and GPT-3.5-Turbo. As targets, we consider four leading frontier models: GPT-4o, Gemini-2.5-Flash, Claude-Sonnet-4.5, and GPT-5. Notably, GPT-5 is trained with a safe-completion mechanism, and Claude-Sonnet-4.5 is widely regarded for its robust safeguards. We report the total attack success rate, along with its decomposition into direct success (the model directly outputs harmful content) and para success (harmful content is available only via alternatives). Results are presented in Table 1 and Table 2.

5.1. Results and Discussion

Our results reveal that intention-deception is highly effective across both textual and multimodal settings. As shown in Table 1, our method consistently achieves significantly higher total attack success rates compared to prior multi-turn jailbreak strategies. On easier targets such as GPT-4o and Gemini-2.5-Flash, our approach achieves near-saturation performance, with total success rates often exceeding 90%. More importantly, on challenging targets like Claude-Sonnet-4.5 and GPT-5, where existing methods almost entirely fail, our method maintains substantial success.

Notably, our attack is the only one that reliably elicits para-jailbreaking: scenarios where the model refuses the harmful prompt but still provides exploitable alternative responses. These successes are especially prominent on robust models such as GPT-5 and Claude-Sonnet-4.5, where direct completions are heavily suppressed. This supports our theoretical insight that steering the conversation into high-risk refusal-with-alternative states increases the likelihood of indirect failures, even when the model's top-level behavior remains aligned. This shows that indirect leakage is the key weakness of safe-completion models.

Table 1. Attack success rates (SR) on AdvBench(Text) and ClearHarm. We report **Total SR**, along with its breakdown into **Direct SR** (model directly outputs harmful content) and **Para SR** (harmful content appears in alternatives only).

Attack Method	Attack Model	Victim Model	AdvBench (Text)			ClearHarm		
			Total SR	Direct SR	Para SR	Total SR	Direct SR	Para SR
Chain of Attack [21]	Qwen-Plus	GPT-4o	0.34	0.34	0.00	0.22	0.22	0.00
		Gemini-2.5-Flash	0.44	0.44	0.00	0.28	0.28	0.00
		Claude-Sonnet-4.5	0.00	0.00	0.00	0.00	0.00	0.00
		GPT-5	0.02	0.00	0.02	0.01	0.00	0.01
	GPT-3.5-Turbo	GPT-4o	0.00	0.00	0.00	0.01	0.01	0.00
		Gemini-2.5-Flash	0.00	0.00	0.00	0.00	0.00	0.00
		Claude-Sonnet-4.5	0.00	0.00	0.00	0.00	0.00	0.00
		GPT-5	0.00	0.00	0.00	0.00	0.00	0.00
Crescendo [16]	Qwen-Plus	GPT-4o	0.43	0.43	0.00	0.34	0.34	0.00
		Gemini-2.5-Flash	0.35	0.35	0.00	0.44	0.44	0.00
		Claude-Sonnet-4.5	0.00	0.00	0.00	0.00	0.00	0.00
		GPT-5	0.02	0.00	0.02	0.02	0.00	0.02
	GPT-3.5-Turbo	GPT-4o	0.31	0.31	0.00	0.33	0.33	0.00
		Gemini-2.5-Flash	0.43	0.43	0.00	0.38	0.38	0.00
		Claude-Sonnet-4.5	0.00	0.00	0.00	0.00	0.00	0.00
		GPT-5	0.00	0.00	0.00	0.03	0.00	0.03
Ours	Qwen-Plus	GPT-4o	0.96	0.96	0.00	0.86	0.86	0.00
		Gemini-2.5-Flash	0.98	0.98	0.00	0.89	0.89	0.00
		Claude-Sonnet-4.5	0.59	0.25	0.34	0.53	0.21	0.32
		GPT-5	0.63	0.12	0.51	0.63	0.11	0.52
	GPT-3.5-Turbo	GPT-4o	0.87	0.87	0.00	0.76	0.76	0.00
		Gemini-2.5-Flash	0.91	0.91	0.00	0.77	0.77	0.00
		Claude-Sonnet-4.5	0.36	0.23	0.13	0.31	0.14	0.17
		GPT-5	0.79	0.19	0.60	0.52	0.09	0.43

Table 2. Attack results on AdvBench-Vision. See Table 1 for details of the evaluation metrics.

Attack Method	Attack Model	Victim Model	Total SR	Direct SR	Para SR
Chain of Attack [21]	Qwen-Plus	GPT-4o	0.36	0.36	0.00
		Gemini-2.5-Flash	0.42	0.42	0.00
		Claude-Sonnet-4.5	0.00	0.00	0.00
		GPT-5	0.00	0.00	0.00
Ours	Qwen-Plus	GPT-4o	0.97	0.97	0.00
		Gemini-2.5-Flash	0.98	0.98	0.00
		Claude-Sonnet-4.5	0.65	0.34	0.31
		GPT-5	0.84	0.23	0.61

We also find that these effects hold across attacker models: even the weaker GPT-3.5-Turbo can reliably succeed under our strategy, indicating that the advantage arises from the intent-deception mechanism itself rather than raw attacker strength. This cross-model robustness shows that intent inversion is a structural weakness, not an artifact of a particular attacker. It also underscores a broader implication of our work: current defenses do not meaningfully account for adversarial manipulation of narrative and conversational trajectory, leaving modern safe-completion systems exposed even when facing comparatively weak adversaries.

Together, the results validate our formal claims: steering conversations into intention, consistent but semantically dangerous regions, without violating any surface rules, can subvert even the most safety-conscious LLMs.

6. Conclusion and Future Work

This work shows that even the most advanced LVLMs remain vulnerable under multi-turn, intention deception attacks. By exploiting safe-completion behavior and the model's drive for conversational consistency, our method reliably steers frontier systems toward harmful outputs, including in highly protected domains such as biological threats. Beyond direct jailbreaks, we uncovered para-jailbreaking, a previously unrecognized class of failure where the model refuses the explicit harmful request but still leaks harmful auxiliary information. This indirect channel is subtle, harder to detect, and likely widespread across safe-completion–based defenses.

Our findings indicate that current safety mechanisms do not adequately anticipate adversaries who manipulate intent over multiple turns, nor do they capture the full space of harmful disclosures. Para-jailbreaking, in particular, demands new evaluation protocols and dedicated mitigation strategies. Looking forward, we aim to develop principled defenses that address both direct and para-jailbreaking, and to establish a more comprehensive understanding of trust dynamics in multi-turn interactions. Robust safety for LVLMs will require systems that reason about intent, context, and consistency in a deeper and more resilient way.

Acknowledgments

This work has been funded in part by the Army Research Laboratory (ARL) award W911QX-24-F-0049, DARPA award FA8750-23-2-1015, ONR award N00014-23-1-2840, and ONR MURI grant N00014-25-1-2116.

References

[1] Giandomenico Cornacchia, Giulio Zizzo, Kieran Fraser, Muhammad Zaid Hameed, Ambrish Rawat, and Mark Purcell. Moje: Mixture of jailbreak experts, naive tabular classifiers as guard for prompt attacks. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 304–315, 2024. 3

[2] Oskar Hollinsworth, Ian McKenzie, Tom Tseng, and Adam Gleave. Clearharm: A more challenging jailbreak dataset. Dataset released by FAR AI, 2025. Accessed via Hugging Face dataset "AlignmentResearch/ClearHarm", URL: https://huggingface.co/datasets/AlignmentResearch/ClearHarm. 7, 6

[3] Xiaojun Jia, Tianyu Pang, Chao Du, Yihao Huang, Jindong Gu, Yang Liu, Xiaochun Cao, and Min Lin. Improved techniques for optimization-based jailbreaking on large language models. *arXiv preprint arXiv:2405.21018*, 2024. 2

[4] Madhur Jindal and Saurabh Deshpande. Reveal: Multi-turn evaluation of image-input harms for vision llm. *arXiv preprint arXiv:2505.04673*, 2025. 3

[5] Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In *European Conference on Computer Vision*, pages 174–189. Springer, 2024. 2

[6] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[7] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, et al. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023. 2

[8] Yue Liu, Xiaoxin He, Miao Xiong, Jinlan Fu, Shumin Deng, and Bryan Hooi. Flipattack: Jailbreak llms via flipping. *arXiv preprint arXiv:2410.02832*, 2024. 2

[9] Weikai Lu, Ziqian Zeng, Jianwei Wang, Zhengdong Lu, Zelin Chen, Huiping Zhuang, and Cen Chen. Eraser: Jailbreaking defense in large language models via unlearning harmful knowledge. *arXiv preprint arXiv:2404.05880*, 2024. 3

[10] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. *arXiv preprint arXiv:2403.09766*, 2024. 2

[11] Yichuan Mo, Yuji Wang, Zeming Wei, and Yisen Wang. Fight back against jailbreaking via prompt adversarial tuning. *Advances in Neural Information Processing Systems*, 37:64242–64272, 2024. 3

[12] Yu Peng, Zewen Long, Fangming Dong, Congyi Li, Shu Wu, and Kai Chen. Playing language game with llms leads to jailbreaking. *arXiv preprint arXiv:2411.12762*, 2024. 2

[13] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 21527–21536, 2024. 2

[14] Qibing Ren, Chang Gao, Jing Shao, Junchi Yan, Xin Tan, Wai Lam, and Lizhuang Ma. Codeattack: Revealing safety generalization challenges of large language models via code completion. *arXiv preprint arXiv:2403.07865*, 2024. 2

[15] Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. Llms know their vulnerabilities: Uncover safety gaps through natural distribution shifts. *arXiv preprint arXiv:2410.10700*, 2024. 2

[16] Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo {Multi-Turn}{LLM} jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 2421–2440, 2025. 2, 7, 8

[17] Rylan Schaeffer, Dan Valentine, Luke Bailey, James Chua, Cristobal Eyzaguirre, Zane Durante, Joe Benton, Brando Miranda, Henry Sleight, John Hughes, et al. Failures to find transferable image jailbreaks between vision-language models. *arXiv preprint arXiv:2407.15211*, 2024. 2

[18] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685, 2024. 2

[19] Cheng Wang, Yue Liu, Baolong Li, Duzhen Zhang, Zhongzhi Li, and Junfeng Fang. Safety in large reasoning models: A survey. *arXiv preprint arXiv:2504.17704*, 2025. 3

[20] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. *arXiv preprint arXiv:2402.08983*, 2024. 3

[21] Xikang Yang, Xuehai Tang, Songlin Hu, and Jizhong Han. Chain of attack: a semantic-driven contextual multi-turn attacker for llm. *arXiv preprint arXiv:2405.05610*, 2024. 2, 7, 8

[22] Yuan Yuan, Tina Sriskandarajah, Anna-Luisa Brakman, Alec Helyar, Alex Beutel, Andrea Vallone, and Saachi Jain. From hard refusals to safe-completions: Toward output-centric safety training. *arXiv preprint arXiv:2508.09224*, 2025. 1, 3

[23] Ziyang Zhang, Qizhen Zhang, and Jakob Foerster. Parden, can you repeat that? defending against jailbreaks via repetition. *arXiv preprint arXiv:2405.07932*, 2024. 3

[24] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. 2, 7, 6