Journal of Information and Communication Technology

Volume 18, Issue 68-69, Spring and Summer 2026, pp. 171-197

A Systematic Review of Trustworthiness, Hallucination, and Safety in Large Vision-Language Models

Delaram Kiani[1], Hanieh Naderi[2*]

[1] PhD Student, School of Intelligent Systems Engineering, College of Interdisciplinary Science and Technology, University of Tehran, Tehran, Iran

[2] Assistant Professor, School of Intelligent Systems Engineering, College of Interdisciplinary Science and Technology, University of Tehran, Tehran, Iran

Received: 18 April 2026, Revised: 21 July 2026, Accepted: 23 August 2026
Paper type: Review

Abstract

Large Vision–Language Models (LVLMs), despite their remarkable progress in multimodal tasks, continue to face two fundamental reliability challenges: the generation of hallucinatory content inconsistent with visual evidence, and vulnerability to adversarial attacks that bypass safety mechanisms. This systematic review, conducted in accordance with the PRISMA 2020 reporting framework, covers studies published from January 2022 to June 2026 and synthesizes 31 eligible studies. It investigates the technical roots of hallucination across multiple levels (object, attribute, relation, and narrative), visual jailbreak mechanisms, and mitigation strategies at both training and inference stages. The findings indicate that over-reliance on linguistic priors is a common underlying factor behind many hallucination errors and safety failures, and that purely text-based alignment is insufficient for multimodal systems. Furthermore, a significant gap exists between current evaluation capabilities and the real-world complexity of vulnerabilities, hindering comprehensive reliability assessment. This review provides an integrated taxonomy of vulnerabilities and outlines a research roadmap, offering a structured foundation for the development of more reliable vision–language models.

Keywords: Vision–Language Models, Reliability, Hallucination, Multimodal Safety, Visual Jailbreak.

* Corresponding Author's email: Hanieh.naderi@ut.ac.ir

دوفصلنامه
فناوری اطلاعات و ارتباطات ایران

سال هجدهم، شماره‌های ۶۸ و ۶۹ ، بهار و تابستان ۱۴۰۵، صفحه ۱۷۱ الی ۱۹۷

مروری نظام‌مند بر قابلیت اعتماد، توهم و ایمنی در مدل‌های بزرگ بینایی-زبان

دلارام کیانی[۱]، حانیه نادری[۲]*

[۱] دانشجوی دکتری دانشکده مهندسی سامانه‌های هوشمند، دانشکدگان علوم و فناوری‌های میان‌رشته‌ای، دانشگاه تهران، تهران، ایران

[۲] استادیار دانشکده مهندسی سامانه‌های هوشمند، دانشکدگان علوم و فناوری‌های میان‌رشته‌ای، دانشگاه تهران، تهران، ایران

تاریخ دریافت: ۱۴۰۵/۰۱/۲۹ تاریخ بازبینی: ۱۴۰۵/۰۴/۳۰ تاریخ پذیرش: ۱۴۰۵/۰۶/۰۱

نوع مقاله: مروری

چکیده

مدل‌های بزرگ بینایی-زبان با وجود پیشرفت‌های چشمگیر در وظایف چندوجهی، همچنان با دو چالش بنیادین قابلیت اعتماد روبه‌رو هستند: تولید محتوای توهمی ناسازگار با شواهد بصری، و آسیب‌پذیری در برابر حملات خصمانه که سازوکارهای ایمنی را دور می‌زنند. این مرور نظام‌مند، با پیروی از چارچوب گزارش‌دهی پریزما ۲۰۲۰، مطالعات منتشرشده از ژانویه ۲۰۲۲ تا ژوئن ۲۰۲۶ را پوشش می‌دهد و ۳۱ مطالعه واجد شرایط را ترکیب و تحلیل می‌کند. این مطالعه ریشه‌های فنی توهم در سطوح مختلف شامل شیء، ویژگی، رابطه و روایت، سازوکارهای گریز امنیتی بصری و راهبردهای کاهش در زمان آموزش و استنتاج را بررسی می‌کند. یافته‌های این مطالعه نشان می‌دهد که اتکای بیش از حد به پیشین‌های زبانی، ریشه مشترک بسیاری از خطاهای توهمی و شکست‌های ایمنی است و هم‌ترازی صرفاً متنی برای سیستم‌های چندوجهی کافی نیست. همچنین، شکاف معناداری میان توانمندی‌های ارزیابی فعلی و پیچیدگی واقعی آسیب‌پذیری‌ها وجود دارد که مانع سنجش جامع قابلیت اعتماد می‌شود. این مرور با ارائه رده‌بندی یکپارچه از آسیب‌پذیری‌ها و نقشه راه پژوهشی، چارچوبی برای توسعه مدل‌های قابل اعتمادتر فراهم می‌آورد.

کلیدواژگان: مدل‌های بینایی-زبان، قابلیت اعتماد؛ توهم، ایمنی چندوجهی، گریز امنیتی بصری.

* رایانامهٔ نویسنده مسؤول: Hanieh.naderi@ut.ac.ir

۱- مقدمه

در سال‌های اخیر، مدل‌های بزرگ بینایی-زبان[1] با تلفیق رمزگذارهای بصری و مدل‌های زبانی بزرگ، به سطحی از توانایی در درک و تولید پاسخ‌های مبتنی بر تصویر رسیده‌اند که آن‌ها را برای طیفی از کاربردهای واقعی جذاب کرده است[1].

این جذابیت به‌ویژه در حوزه‌های حساس مانند تحلیل تصویر در تصمیم‌گیری‌های پرریسک، سامانه‌های تعاملی، و سناریوهایی که «توصیف/استدلال درباره تصویر» مستقیماً وارد چرخه تصمیم می‌شود، پررنگ‌تر است[2].

با وجود این موفقیت‌ها، شواهد پژوهشی نشان می‌دهد که عملکرد خوب در سنجه‌های[2] رایج، لزوماً به معنای قابلیت اعتماد در دنیای واقعی نیست و مدل‌ها در موقعیت‌های طبیعی، پرسش‌های چندمرحله‌ای، یا تولید آزاد می‌توانند دچار خطاهای نظام‌مند شوند [1].

یکی از نمودهای برجسته این وضعیت، «توهم» است؛ یعنی تولید محتوایی که با شواهد بصری تصویر همخوان نیست و می‌تواند شامل نسبت‌دادن اشیاء، ویژگی‌ها یا روابطی باشد که در تصویر وجود ندارند[2].

اهمیت این مسئله زمانی حیاتی‌تر می‌شود که خروجی مدل در محیط‌های حساس یا تخصصی به‌عنوان مبنای تصمیم تلقی شود، زیرا حتی خطاهای کوچک می‌توانند پیامدهای بزرگ ایجاد کنند[3].

علاوه بر توهم، پژوهش‌های جدید نشان داده‌اند که افزودن ماهیتی بصری می‌تواند مسیرهای تازه‌ای برای دور زدن سیاست‌های ایمنی باز کند و مدل‌های چندوجهی را در معرض «گریز امنیتی[3] بصری» قرار دهد[5] ,[4]. برای نمونه، معیار ایمنی چندوجهی[4] نشان می‌دهد که حتی تصاویر مرتبط با پرسش می‌توانند سازوکارهای ایمنی را تضعیف کنند، حتی وقتی مدل زبانی زیربنایی قبلاً هم‌ترازی ایمنی[5] شده باشد[4].

همچنین گریز امنیتی وی-۲۸هزار[6] به‌طور نظام‌مند نشان می‌دهد تکنیک‌های گریز امنیتی موفق در مدل‌های صرفاً متنی می‌توانند به فضای چندوجهی منتقل شوند و نرخ موفقیت حمله[7] قابل توجهی ایجاد کنند [5].

پژوهش‌های بعدی نیز نشان می‌دهند که حملات مبتنی بر «حواس‌پرت‌سازی چندتصویری» یا طراحی ورودی‌های مرکب می‌توانند با پراکنده‌کردن توجه مدل، سازوکارهای دفاعی را دور بزنند [6].

این هم‌نشینیِ توهم و شکست ایمنی، به یک مسئله مرکزی منجر می‌شود که می‌توان آن را «شکاف اعتماد» نامید؛ یعنی فاصله بین موفقیت در ارزیابی‌های آزمایشگاهی و قابلیت اتکای واقعی در استقرار [4] ,[2]. برای تحلیل دقیق‌تر اعتمادپذیری در مدل‌های بینایی-زبان، تفکیک مفهومی میان دو معیار بنیادین ضروری است [7].

وفاداری به سازگاری درونی[8] اشاره دارد و این پرسش را مطرح می‌کند که آیا متن تولید شده توسط مدل، به‌طور دقیق محتوای موجود در تصویر ورودی را منعکس می‌کند یا خیر. در ادبیات ارزیابی توهم، وفاداری معمولاً در قالب خطاهای شیء، ویژگی و رابطه صورت‌بندی می‌شود و حتی رده‌های پیچیده‌تری مانند «توهم رویداد» نیز برای پوشش خطاهای روایی معرفی شده‌اند [7].

واقعیت‌محوری به سازگاری بیرونی[9] مربوط می‌شود و بررسی می‌کند که آیا خروجی مدل با حقایق معتبر جهان واقعی سازگار است یا خیر، حتی اگر توصیف ظاهری تصویر دقیق به نظر برسد [1].

به بیان دیگر، یک مدل ممکن است در سطح توصیف تصویر خطای آشکاری نداشته باشد، اما در تفسیر، استنتاج یا ادعاهای دانش‌محور دچار خطا شود که در کاربردهای حساس می‌تواند خطرناک‌تر از خطای توصیفی باشد [3]. این تمایز از آن جهت مهم است که بسیاری از راهبردهای کاهش توهم، عمدتاً وفاداری را هدف می‌گیرند و الزاماً تضمینی برای واقعیت‌محوری ارائه نمی‌دهند. یکی از توضیح‌های ریشه‌ای برای بخشی از شکست‌های وفاداری، پدیده‌ای موسوم به «شکاف ماهیتی» است که به فاصله هندسی میان بازنمایی‌های تصویر و متن در فضای مشترک embedding اشاره دارد [8].

تحلیل لیانگ و همکاران نشان می‌دهد در مدل‌های کنتراستیو چندوجهی (مانند CLIP)، دو ماهیت می‌توانند «در کنار هم» اما نه «به‌صورت واقعاً ادغام‌شده» بازنمایی شوند و همین امر، اتکا به

[1] Large Vision-Language Models
[2] Benchmark
[3] JailBreak
[4] MM-SafetyBench
[5] Safety-aligned
[6] JailBreakV-28K
[7] Attack Success Rate (ASR)
[8] Internal Consistency
[9] External Consistency

سیگنال‌های غیر بصری را در موقعیت‌های مبهم افزایش می‌دهد [8].

در ادامه، شرودی و همکاران نشان می‌دهند شکاف وجهیت و پدیده‌هایی مانند سوگیری شیء می‌توانند از یک محرک مشترک ناشی شوند؛ یعنی عدم‌تعادل اطلاعاتی میان تصاویر و توضیحات متنی در داده‌های آموزشی که به شکل‌گیری الگوهای نامتوازن در یادگیری می‌انجامد [9].

همچنین ادبیات جدید بین «شکاف ماهیتی» و «شکاف قابلیت» تمایز می‌گذارد و نشان می‌دهد حتی اگر هم‌ترازی بازنمایی‌ها بهبود یابد، ممکن است پیش‌بینی عملکرد واقعی روی داده یا وظیفه هدف همچنان دشوار بماند [10].

این مرور نظام‌مند با تکیه بر چارچوب گزارش‌دهی پریزما ۲۰۲۰[1] طراحی شده است تا ادبیات مرتبط با توهم، اعتمادپذیری و ایمنی در مدل‌های بزرگ بینایی-زبان را به‌صورت قابل تکرار و قابل دفاع گردآوری و تحلیل کند. سؤال نخست این است که ریشه‌های فنی توهم در VLMها چیست و نقش داده، معماری، و پیشین‌های زبانی در بروز خطاهای شیء/ویژگی/رابطه چگونه قابل توضیح است. سؤال دوم بررسی می‌کند حملات خصمانه بصری چگونه می‌توانند سازوکارهای ایمنی را دور بزنند و چرا هم‌ترازی ایمنیِ صرفاً متنی برای سیستم‌های چندوجهی کافی نیست.

سؤال سوم به اثربخشی راهبردهای کاهش در مرحله آموزش و مرحله استنتاج می‌پردازد و مشخصاً می‌پرسد روش‌های مبتنی بر یادگیری ترجیح (مانند بهینه‌سازی ترجیح مستقیم و نسخه‌های هدایت‌شده با سیگنال بصری) چه میزان می‌توانند اتکای بیش‌ازحد به پیشین‌های زبانی را کاهش دهند. سؤال چهارم به «شکاف ارزیابی» مربوط است و بررسی می‌کند چرا مجموعه سنجه‌ها و متریک‌های فعلی هنوز پوشش یکپارچه‌ای از انواع توهم (از ساده تا روایی) و هم‌زمان تهدیدهای ایمنی ارائه نمی‌کنند. در ادامه، مقاله ابتدا روش‌شناسی مرور نظام‌مند و پروتکل انتخاب و ارزیابی مطالعات را بر اساس پریزما تشریح می‌کند تا مسیر گردآوری شواهد شفاف و تکرارپذیر باشد. سپس یک رده‌بندی[2] تحلیلی از آسیب‌پذیری‌ها ارائه می‌شود که هم انواع توهم (شیء، ویژگی، رابطه و موارد پیچیده‌تر) را پوشش می‌دهد و هم تهدیدهای ایمنی مانند گریز امنیتی بصری را در یک چارچوب واحد صورت‌بندی می‌کند. در گام بعد، اکوسیستم ارزیابی شامل سنجه‌های تشخیصی و ایمنی و شکاف‌های سنجش بررسی می‌شود تا روشن شود چرا «اندازه‌گیری» در این حوزه خود یک چالش بنیادین است. پس از آن، راهبردهای کاهش در زمان آموزش و استنتاج مقایسه می‌شوند تا مبادله‌های هزینه، کارایی و قابلیت استقرار در شرایط واقعی مشخص شود. در پایان، مقاله با بحث درباره چرایی تداوم شکاف اعتماد و مسیرهای آینده مانند گسترش سطح حمله در ماهیت‌های پیچیده‌تر (از جمله ویدئو) جمع‌بندی می‌شود تا نقشه راه پژوهشی روشن‌تری ارائه گردد. برای ایجاد درکی یکسان از اصطلاحات اصلی مورد استفاده در این مرور، مفاهیم و تعاریف کلیدی در جدول ۱ ارائه شده‌اند.

همان‌طور که جدول ۲ نشان می‌دهد، مرورهای پیشین عمدتاً بر یکی از دو محور «توهم» یا «ایمنی» تمرکز داشته‌اند و کمتر به تعامل این دو آسیب‌پذیری و ریشه‌های مشترک آن‌ها پرداخته‌اند. نوآوری اصلی این مرور، یکپارچه‌سازی هر دو محور در یک چارچوب تحلیلی واحد، ارائه رده‌بندی جامع از انواع توهم (از سطح شیء تا روایت)، و بررسی هم‌زمان تهدیدهای ایمنی مانند گریز امنیتی بصری است. همچنین، این مطالعه نخستین مروری است که با پروتکل پریزما ۲۰۲۰ در این حوزه خاص انجام شده و قابلیت تکرارپذیری و شفافیت روش‌شناختی را تضمین می‌کند.

جدول ۱. مفاهیم و تعاریف کلیدی در این مرور

مفهوم (FA/EN)	تعریف دقیق	تمرکز ارزیابی	مثال کوتاه	منابع/سنجه‌های نماینده
شکاف اعتماد / Trust Gap	شکاف اعتماد به فاصله بین عملکرد مطلوب مدل در ارزیابی‌های استاندارد و «قابلیت اتکا» آن در سناریوهای واقعی (از نظر صحت/وفاداری و نیز ایمنی در برابر ورودی‌های مسئله‌ساز) اشاره دارد.	سنجش شکاف اعتماد معمولاً با مقایسه نتایج سنجه‌ی با رفتار مدل در تولید آزاد، مکالمه چندمرحله‌ای، یا ورودی‌های دستکاری‌شده انجام می‌شود.	مدلی که روی QA تصویری نمره بالا می‌گیرد، اما در توضیح آزاد تصویر جزئیات ساختگی اضافه می‌کند، نمونه‌ای از شکاف اعتماد را نشان می‌دهد.	POPE [۲]; معیار ایمنی چندوجهی [4]
قابلیت اعتماد / Trustworthiness	قابلیت اعتماد یک مفهوم چتری است که کیفیت خروجی را از چند محور شامل وفاداری به تصویر، صحت ادعاها و پایبندی به ایمنی در شرایط عادی و خصمانه توصیف می‌کند.	ارزیابی قابلیت اعتماد نیازمند سنجش هم‌زمان خطاهای توهم و شکست‌های ایمنی، معمولاً با ترکیب سنجه‌های توهم و سنجه‌های امنیت/ایمنی است.	در یک سامانه حساس، خروجی باید هم «درست درباره تصویر» باشد و هم «به سیاست‌های ایمنی پایبند» بماند تا قابل اعتماد تلقی شود.	POPE[2]; JailBreakV-28K [۵]; معیار ایمنی چندوجهی [4]

[1] PRISMA 2020

[2] Benchmark

وفاداری / Faithfulness	وفاداری به سازگاری درونی خروجی با شواهد بصری اشاره دارد و می‌پرسد آیا مدل به آنچه در تصویر قابل مشاهده/استنتاج است پایبند مانده یا محتوای بی‌پشتوانه افزوده است.	تمرکز ارزیابی وفاداری روی هم‌راستایی تصویر-متن و تشخیص ناسازگاری‌هایی مثل خطاهای شیء/ویژگی/رابطه است.	اگر تصویر فقط «مبل» دارد اما مدل «سگ روی مبل» را توصیف کند، وفاداری نقض شده است.	Hal-Eval[7]; POPE [2]
واقعیت‌محوری / Factuality	واقعیت‌محوری به سازگاری بیرونی خروجی با حقایق معتبر جهان واقعی اشاره دارد و حتی در صورت توصیف درست تصویر، ادعاهای دانشی نادرست را خطا محسوب می‌کند.	ارزیابی واقعیت‌محوری معمولاً نیازمند راستی‌آزمایی با منابع بیرونی، مدل‌های داور، یا پروتکل‌هایی است که ادعا را جدا از صرفِ مشاهده تصویر می‌سنجند.	مدل ممکن است تصویر دارو را درست توصیف کند اما درباره کاربرد/دوز آن ادعای غلط بدهد و واقعیت‌محوری پایین داشته باشد.	Hal-Eval [7].
توهم / Hallucination	توهم به تولید ادعاهای ناسازگار با تصویر اطلاق می‌شود و در مدل‌های بزرگ بینایی-زبان معمولاً به صورت توهم شیء، ویژگی، رابطه و حتی «توهم رویداد» صورت‌بندی می‌شود.	تمرکز ارزیابی توهم روی شناسایی خطاهای ریزدانه و سنجش نرخ/شدت آن‌ها در سناریوهای پرسش‌پاسخ و تولید آزاد است.	گفتن «سه نفر در تصویر هستند» وقتی دو نفر حضور دارند، نمونه‌ای از توهم ویژگی (تعداد) است.	POPE [2]; Hal-Eval [7], [11].
شکاف ماهیتی / Modality Gap	شکاف ماهیتی پدیده‌ای هندسی در فضای بازنمایی مشترک است که در آن بردارهای تصویر و متن «با فاصله» در کنار هم قرار می‌گیرند و ادغام کامل رخ نمی‌دهد.	تمرکز ارزیابی شکاف ماهیتی معمولاً بر تحلیلembedding ها (مثلاً جدایی خوشه‌های تصویر و متن) و مطالعه اثر آن بر عملکرد و تعمیم مدل است.	وقتی مدل در ابهام‌ها به پیش‌فرض زبانی تکیه می‌کند و تصویر را کم‌اثر می‌بیند، شکاف ماهیتی می‌تواند یکی از توضیح‌های محتمل باشد.	Mind the Gap [8], [12], [13]
گریز امنیتی بصری / Visual Jailbreaking	گریز امنیتی بصری به استفاده از ورودی تصویر یا ترکیب تصویر-متن برای دور زدن هم‌ترازی ایمنی و وادارکردن مدل به تولید خروجی ممنوع/مضر گفته می‌شود.	تمرکز ارزیابی گریز امنیتی بصری روی اندازه‌گیری میزان نفوذ حمله و سنجش حساسیت مدل به سناریوهای مختلف دستکاری (مثل تصاویر مرتبطِ مخرب یا حواس‌پرت‌سازی چندتصویری) است.	ترکیب یک پرسش مضر با تصویری که ظاهراً «مرتبط» است و باعث تولید پاسخ ممنوع می‌شود، نمونه‌ای از گریز امنیتی بصری است.	معیار ایمنی چندوجهی [14]; JailBreakV-28K [5]; CS-DJ [6].
همترازی ایمنی / Safety Alignment	همترازی ایمنی به مجموعه روش‌هایی گفته می‌شود که هدفشان تقویت امتناع/محدودسازی تولید محتوای پرریسک و افزایش مقاومت مدل در برابر ورودی‌های مسئله‌ساز چندوجهی است.	تمرکز ارزیابی همترازی ایمنی بر کاهش خروجی‌های ناقض سیاست و بهبود امتیاز مدل روی سنجه‌های ایمنی و ردتیمینگ است .	افزودن ماژول‌ها یا اهداف آموزشی که باعث می‌شود مدل در مواجهه با تصویر تحریک‌کننده، پاسخ ایمن بدهد، نمونه‌ای از همترازی ایمنی است.	Safety Alignment for Vision Language Models [4].
نرخ موفقیت حمله / Attack Success Rate (ASR)	نرخ موفقیت حمله درصد نمونه‌های حمله‌ای است که در آن‌ها مدل به خروجی ناایمن/ناقض سیاست منجر می‌شود و معیار رایج گزارش‌گری در ارزیابی گریز امنیتی است.	تمرکز ارزیابی با نرخ موفقیت حمله روی مقایسه مدل‌ها و دفاع‌ها در سناریوهای حمله مختلف و تحلیل شرایطی است که نرخ موفقیت حمله را بالا یا پایین می‌برد.	اگر از ۱۰۰ ورودی حمله، مدل در ۶۰ مورد پاسخ ممنوع تولید کند، نرخ موفقیت حمله برابر ۶۰٪ است.	JailBreakV-28K [5]; CS-DJ [6].

جدول ۲. مقایسه مرورهای نظام‌مند موجود در حوزه مدل‌های بزرگ بینایی-زبان

ویژگی /مرور	مرور توهم LLM [16]	مرور توهم MLLM [17]	Rawte و همکاران (2023)	مرور ارزیابی LVLM [3]	این مرور
تمرکز اصلی	توهم در مدل‌های زبانی بزرگ	ایمنی مدل‌های چندوجهی	توهم در تولید متن	ارزیابی مدل‌های بینایی-زبان	توهم + ایمنی + تحلیل ریشه‌ای (یکپارچه)
پوشش توهم شیء	✓	✗	✓	✓	✓
پوشش توهم ویژگی/رابطه	محدود	✗	محدود	✓	✓
پوشش توهم روایی/رویداد	✗	✗	✗	محدود	✓
تحلیل گریز امنیتی بصری	✗	✓	✗	✗	✓
تحلیل ریشه‌ای)پیشین زبانی)	محدود	✗	✗	✗	✓
اکوسیستم ارزیابی یکپارچه	✗	✗	✗	✓	✓
راهبردهای کاهش (آموزش/استنتاج)	محدود	محدود	✗	محدود	✓
چارچوب گزارش‌دهی	روایی	روایی	روایی	روایی	پریزما ۲۰۲۰
دامنه زمانی	تا 2023	تا 2024	تا 2023	تا 2024	2023-2025

۱-۱- مرور مطالعات پیشین و شکاف‌های پژوهشی

در سال‌های اخیر، بخش قابل توجهی از پژوهش‌های مرتبط با مدل‌های بزرگ بینایی-زبان بر مسئله توهم و ریشه‌های ایجاد آن متمرکز بوده‌اند [1], [2], [3], [16], [17] مطالعات اولیه بیشتر بر شناسایی و اندازه‌گیری خطاهای شیء و ویژگی تمرکز داشتند و تلاش می‌کردند میزان ناسازگاری میان خروجی مدل و محتوای تصویر را ارزیابی کنند[2]. در ادامه، پژوهش‌های جدیدتر به تحلیل ریشه‌های این خطاها از منظر سوگیری‌های زبانی، عدم‌تعادل داده‌های آموزشی، و ضعف هم‌ترازی میان بازنمایی‌های تصویری و متنی پرداختند [12], [9], [18]. همچنین، مرورها و سنجه‌های جدید نشان داده‌اند که توهم تنها به خطاهای ساده محدود نیست و در سناریوهای پیچیده‌تری مانند استدلال چندمرحله‌ای، تولید آزاد، و تعامل چندنوبتی می‌تواند به شکل‌های روایی و زمینه‌ای نیز ظاهر شود [1], [3] , [19], [20].

همزمان، گروه دیگری از پژوهش‌ها بر طراحی سنجه‌ها و چارچوب‌های ارزیابی توهم تمرکز کرده‌اند. این مطالعات نشان داده‌اند که ارزیابی توهم به‌شدت به نوع پرسش، قالب پاسخ، و سناریوی آزمون وابسته است و ممکن است یک مدل در یک سنجه عملکرد مطلوبی داشته باشد، اما در محیط‌های پیچیده‌تر همچنان دچار خطا شود [7], [8]. در ادامه، برخی پژوهش‌ها تلاش کردند ارزیابی را از تصاویر منفرد به سناریوهای دشوارتری مانند تعامل چندنوبتی، استدلال چندمرحله‌ای، و تحلیل تغییرات زمینه‌ای گسترش دهند[8], [19], [21] همچنین، پژوهش‌های جدیدتر نشان داده‌اند که تفاوت در تعریف معیارها، سطح جزئیات ارزیابی، و نحوه طراحی پروتکل‌های آزمون می‌تواند مقایسه مستقیم مدل‌ها را دشوار کند و حتی به نتایج ناپایدار یا سوگیرانه منجر شود [20]. این یافته‌ها نشان می‌دهد که اکوسیستم ارزیابی فعلی هنوز فاقد یک چارچوب جامع و یکپارچه برای سنجش هم‌زمان توهم و قابلیت اعتماد در شرایط واقعی است.

در کنار مسئله توهم، ادبیات جدید توجه گسترده‌ای به تهدیدهای ایمنی و حملات گریز امنیتی در مدل‌های بزرگ بینایی-زبان نشان داده است [4], [22], [14] مطالعات اخیر نشان می‌دهند که بسیاری از سازوکارهای هم‌ترازی ایمنی که برای مدل‌های صرفاً متنی طراحی شده‌اند، در محیط‌های چندوجهی پایداری کافی ندارند [4] [15]. برای نمونه، JailBreakV نشان می‌دهد که تکنیک‌های موفق گریز امنیتی در مدل‌های متنی می‌توانند به فضای چندوجهی منتقل شوند و نرخ موفقیت حمله قابل توجهی ایجاد کنند[5]. همچنین، پژوهش‌های جدید در حوزه حملات ترکیبی و حواس‌پرت‌سازی بصری نشان داده‌اند که طراحی ورودی‌های مرکب می‌تواند سازوکارهای دفاعی مدل‌های چندوجهی را دور بزند[6], [23]. در ادامه، پژوهش‌های جدیدتر با گسترش حملات به ماهیت ویدئویی و فرایندهای استدلال تصویری، سطح جدیدی از آسیب‌پذیری‌ها را آشکار کرده‌اند[24], [25]. همزمان، برخی مطالعات نیز راهبردهای دفاعی جدیدی برای کاهش آسیب‌پذیری مدل‌ها در برابر حملات گریز امنیتی پیشنهاد کرده‌اند[26], [27].

با وجود گسترش سریع ادبیات، همچنان چند شکاف اساسی در پژوهش‌های موجود مشاهده می‌شود. نخست آنکه بسیاری از مطالعات، توهم و ایمنی را به‌صورت جداگانه بررسی کرده‌اند و ارتباط میان این دو آسیب‌پذیری کمتر به‌صورت یکپارچه تحلیل شده است [4], [5], [16] دوم آنکه بخش قابل توجهی از پژوهش‌ها روی تصاویر منفرد متمرکز بوده و سناریوهای پیچیده‌تری مانند ویدئو، تعامل چندمرحله‌ای[28]، و استدلال زمینه‌ای هنوز پوشش محدودی دارند [24]. همچنین، نبود چارچوب‌های ارزیابی یکپارچه و تفاوت در تعریف معیارها و پروتکل‌های سنجش، مقایسه مستقیم مدل‌ها و تحلیل جامع قابلیت اعتماد آن‌ها را دشوار کرده است[7], [20]. ازاین‌رو، نیاز به یک مرور نظام‌مند که هم مسئله توهم و هم تهدیدهای ایمنی را در یک چارچوب تحلیلی واحد بررسی کند، همچنان احساس می‌شود.

۲- روش‌شناسی

این مطالعه به‌صورت یک مرور نظام‌مند طراحی و گزارش شده است تا فرآیند شناسایی، غربال‌گری، ارزیابی و ترکیب شواهد به‌صورت شفاف و قابل تکرار انجام شود. منطق کلی مرور بر مبنای چارچوب PRISMA 2020 شامل چهار گام شناسایی، غربال‌گری، ارزیابی صلاحیت و گنجاندن نهایی تنظیم شد. دامنه زمانی این مرور، مطالعات منتشرشده از ژانویه ۲۰۲۲ تا ۳۰ ژوئن ۲۰۲۶ را پوشش می‌دهد. جست‌وجوی نظام‌مند اولیه در تاریخ ۳ فوریه ۲۰۲۶ انجام شد. برای اطمینان از به‌روز بودن مرور، جست‌وجوی مکمل در تاریخ ۳۰ ژوئن ۲۰۲۶ انجام گرفت و این تاریخ به‌عنوان آخرین تاریخ جست‌وجو در نظر گرفته شد. پس از اعمال معیارهای ورود و خروج، ۳۱ مطالعه واجد شرایط در سنتز نهایی گنجانده شدند. مطالعات واردشده به سنتز نهایی بین سال‌های ۲۰۲۲ تا ۲۰۲۵ منتشر شده بودند و هیچ‌یک از رکوردهای منتشرشده در سال ۲۰۲۶ معیارهای ورود را احراز نکردند. منابع هدف شامل arXiv، ACL Anthology، مخازن رسمی کنفرانس‌ها/ناشران کلیدی مانند NeurIPS، CVPR، ICLR، EMNLP، ACM Digital Library و IEEE Xplore، و

کنترل ارجاعات رو به عقب/رو به جلو بود. اجزای جست‌وجو شامل محل جست‌وجو، بازه زمانی، کوئری‌ها، ترکیب منطقی کلیدواژه‌ها، رویه ثبت نتایج و حذف موارد تکراری مستندسازی شد. الگوی کلیدواژه‌ها به‌صورت ترکیبی از اصطلاحات مدل و مسئله تعریف شد تا ابعاد توهم، اعتمادپذیری، ایمنی و گریز امنیتی هم‌زمان پوشش داده شود.

("VLM" OR "LVLM" OR "Vision-Language" OR "multimodal LLM") AND ("hallucination" OR "trustworthiness" OR "faithfulness" OR "safety" OR "jailbreak" OR "robustness").

جزئیات پایگاه‌های جست‌وجوشده، تاریخ انجام جست‌وجو و تعداد رکوردهای بازیابی‌شده از هر منبع در جدول ۳ ارائه شده است.

معیارهای ورود و خروج به‌صورت پیشینی تعریف شدند تا انتخاب مطالعات به حداقلِ سلیقه‌گرایی برسد. در این مرور، مطالعاتی پذیرفته شدند که یا مستقیماً توهم در مدل‌های بزرگ بینایی–زبان را اندازه‌گیری/تعریف/تحلیل کرده باشند یا به‌طور تجربی به ایمنی چندوجهی و گریز امنیتی بصری پرداخته باشند یا یک راهبرد کاهش (در زمان آموزش یا استنتاج) برای بهبود وفاداری/ایمنی ارائه داده باشند. مطالعاتی که صرفاً به تولید تصویر (مانند مدل‌های تولیدی تصویر بدون مؤلفه فهم چندوجهی) می‌پرداختند یا صرفاً متن‌محور بودند، از دامنه این مرور کنار گذاشته شدند تا تمرکز مقاله روی مدل‌های بزرگ بینایی-زبان ادراکی/استدلالی و ریسک‌های اعتماد و ایمنی آن‌ها باقی بماند. خلاصه معیارهای ورود و خروج مطالعات در جدول ۴ ارائه شده است.

جدول ۳. پایگاه‌ها، تاریخ‌های جست‌وجو و تعداد رکوردهای بازیابی‌شده پیش از حذف موارد تکراری

تعداد رکورد	تاریخ جست‌وجو اولیه/مکمل	منبع/پایگاه
۲۳۸	۳ فوریه ۲۰۲۶؛ جست‌وجوی مکمل: ۳۰ ژوئن ۲۰۲۶	arXiv
۶۲	۳ فوریه ۲۰۲۶؛ جست‌وجوی مکمل: ۳۰ ژوئن ۲۰۲۶	ACL Anthology / EMNLP
۶۱	۳ فوریه ۲۰۲۶؛ جست‌وجوی مکمل: ۳۰ ژوئن ۲۰۲۶	NeurIPS Proceedings / OpenReview
۵۵	۳ فوریه ۲۰۲۶؛ جست‌وجوی مکمل: ۳۰ ژوئن ۲۰۲۶	CVF Open Access (CVPR/ICCV/ECCV)
۴۸	۳ فوریه ۲۰۲۶؛ جست‌وجوی مکمل: ۳۰ ژوئن ۲۰۲۶	ICLR / OpenReview
۷۸	۳ فوریه ۲۰۲۶؛ جست‌وجوی مکمل: ۳۰ ژوئن ۲۰۲۶	IEEE Xplore ،ACM DL، Springer/ScienceDirect و کنترل ارجاعات
۵۴۲	-	جمع کل

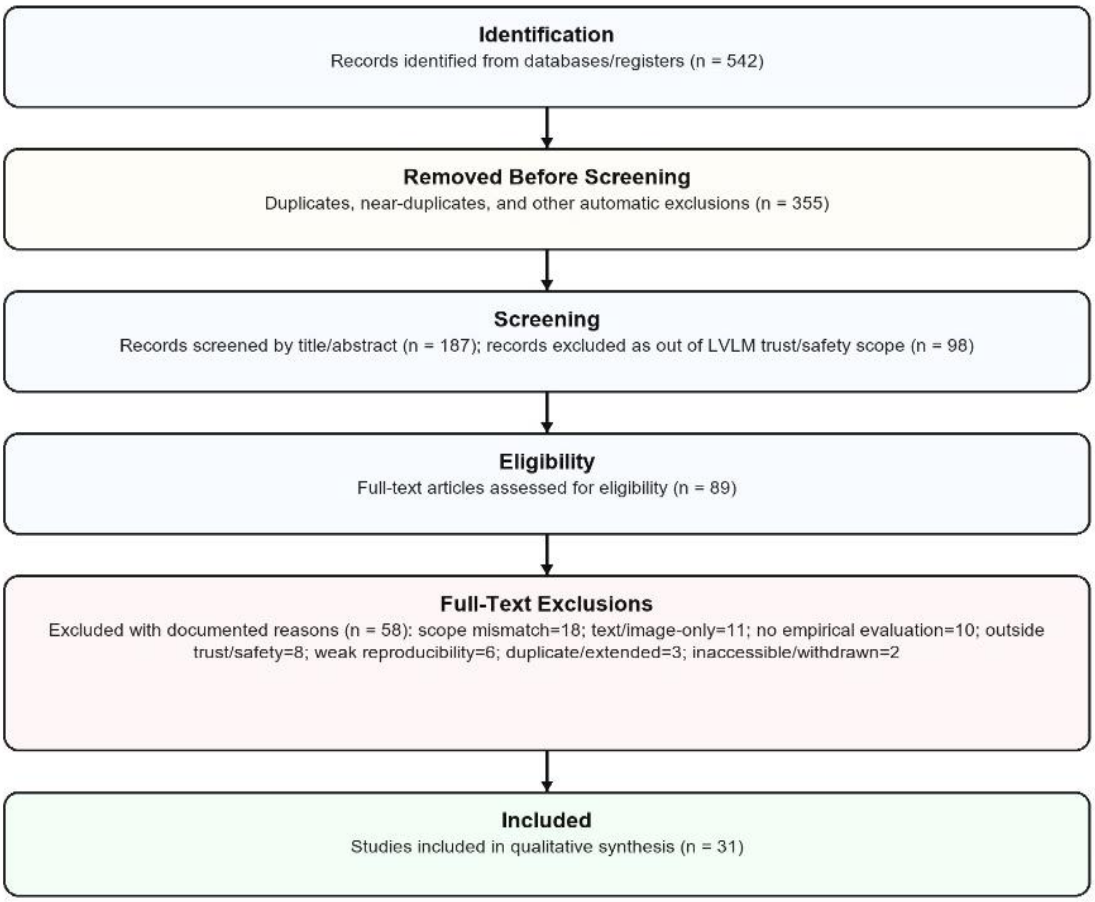
جدول ۴. معیارهای ورود/خروج

دسته	معیارها
Inclusion	(۱) تمرکز مستقیم بر مدل‌های بزرگ بینایی-زبان یا مدل‌های چندوجهی مبتنی بر LLM در وظایف درک/پاسخ‌دهی (نه صرفاً تولید تصویر)؛ (۲) ارائه ارزیابی تجربی یا پروتکل سنجش (سنجه/متریک/مطالعه کاربری/ردتیمینگ)؛ (۳) ارتباط روشن با یکی از محورهای توهم، اعتمادپذیری، ایمنی یا گریز امنیتی بصری؛ (۴) کیفیت انتشار قابل قبول (کنفرانس/ژورنال معتبر یا arXiv با روش و نتایج قابل بازبینی).
Exclusion	(۱) مطالعات صرفاً متن‌محور بدون مؤلفه دیداری؛ (۲) مطالعات صرفاً تولید تصویر بدون مسئله اعتماد/ایمنی در فهم چندوجهی؛ (۳) مقالات موضع‌گیری/دیدگاه بدون نتایج تجربی؛ (۴) آثار فاقد توصیف کافی از داده/مدل/متریک به‌گونه‌ای که بازتولید یا ارزیابی مستقل ممکن نباشد.

برای ارزیابی کیفیت و ریسک سوگیری، از چک‌لیست CASP به‌عنوان چارچوب پرسش‌محور استفاده شد، اما پرسش‌ها برای ماهیت تجربی/مهندسی مطالعات یادگیری ماشین نگاشت عملیاتی شدند. این نگاشت پنج بُعد داشت: (۱) شفافیت تعریف مسئله و مدل تهدید، (۲) کفایت داده‌ها، خط پایه‌ها و سنجه‌ها، (۳) گزارش تنظیمات ارزیابی و امکان بازتولید، (۴) تحلیل حساسیت/مطالعه حذف مؤلفه‌ها یا مقایسه دفاع/حمله، و (۵) شفافیت محدودیت‌ها و گزارش شکست‌ها. دو داور مستقل ابتدا پنج مطالعه را به‌صورت آزمایشی کدگذاری کردند و پس از توافق بر دستورالعمل کدگذاری، همه مطالعات واجد شرایط را با مقیاس سه‌سطحی «کم/متوسط/زیاد» از نظر ریسک سوگیری ارزیابی نمودند. توافق بین‌داوران برای تصمیم ورود/خروج $\kappa=0.82$ و برای کدگذاری کیفیت روش‌شناسی $\kappa=0.79$ به‌دست آمد؛ اختلاف‌ها با گفت‌وگو و در موارد باقی‌مانده با نظر داور سوم حل شد. ریسک سوگیری در سه سطح داده، سنجه/متریک و گزارش‌دهی ثبت شد.

برای هر مطالعه پذیرفته‌شده، استخراج داده با فرم استاندارد انجام شد. فرم استخراج شامل سال و نوع انتشار، خانواده مدل/معماری، نوع مسئله، داده‌ها و سنجه‌ها، متریک‌های کلیدی مانند نرخ توهم

یا ASR/EASR، تنظیمات تهدید، یافته‌های اصلی، محدودیت‌ها و کیفیت گزارش‌دهی بود. دو داور داده‌ها را به‌صورت مستقل استخراج کردند؛ ناسازگاری‌های عددی با بازگشت به متن کامل مقاله و ناسازگاری‌های مفهومی با اجماع داوران اصلاح شد. به دلیل ناهمگونی بالای وظایف، داده‌ها و متریک‌ها، سنتز کمی/متاآنالیز انجام نشد. سنتز روایی در سه گام انجام گرفت: کدگذاری اولیه یافته‌ها بر اساس نوع شکست یا راهکار، خوشه‌بندی کدها در مضامین سطح بالاتر مانند توهم، ایمنی، ریشه‌های شکست و راهبردهای کاهش، و سپس ساخت ماتریس شواهد برای مقایسه مطالعات بر حسب دامنه، مدل تهدید، سنجه و پیامد. نتیجه این فرایند به‌صورت رده‌بندی آسیب‌پذیری‌ها، نقشه اکوسیستم ارزیابی و مقایسه خانواده راهکارهای کاهش گزارش شد. فرایند شناسایی، غربال‌گری، ارزیابی صلاحیت و انتخاب نهایی مطالعات براساس چارچوب PRISMA 2020 در شکل ۱ نشان داده شده است.



شکل ۱. نمودار جریان انتخاب مطالعات

در مرحله شناسایی، جست‌وجوی نظام‌مند در پایگاه‌های داده و رجیستری‌های تعریف‌شده منجر به بازیابی ۵۴۲ رکورد گردید. پیش از ورود به مرحله غربال‌گری، ۳۵۵ رکورد به دلیل تکراری بودن، نسخه‌های بسیار مشابه یا خروج آشکار از دامنه مرور حذف شدند. در مرحله غربال‌گری، ۱۸۷ رکورد بر اساس عنوان و چکیده ارزیابی شدند و ۹۸ رکورد به دلیل عدم ارتباط با مدل‌های بزرگ بینایی-زبان، توهم، اعتمادپذیری، ایمنی یا گریز امنیتی کنار گذاشته شدند. در مرحله ارزیابی صلاحیت، ۸۹ مقاله به‌صورت متن کامل بررسی شدند. از این تعداد، ۵۸ مطالعه بر اساس دلایل ثبت‌شده در جدول ۵ حذف گردیدند و در نهایت ۳۱ مطالعه معیارهای ورود به مرور را احراز نمودند و در سنتز کیفی گنجانده شدند.

جدول ۵. دلایل حذف ۵۸ مقاله در مرحله ارزیابی متن کامل

تعداد	دلیل حذف در ارزیابی متن کامل
۱۱	تمرکز اصلی بر مدل‌های صرفاً متنی یا تولید تصویر، بدون مؤلفه ادراکی LVLM/MLLM
۸	عدم تمرکز مستقیم بر توهم، اعتمادپذیری، ایمنی یا گریز امنیتی بصری
۱۰	نبود ارزیابی تجربی، سنجه، پروتکل آزمون یا داده قابل بررسی
۶	فقدان جزئیات کافی درباره داده، مدل، پرامپت، متریک یا تنظیمات بازتولید
۱۸	خارج بودن از دامنه مدل‌های بزرگ بینایی-زبان یا صرفاً کاربردی/غیرمرتبط بودن
۳	نسخه تکراری، پیش‌چاپ جایگزین‌شده یا نسخه توسعه‌یافته همان کار
۲	عدم دسترسی به متن کامل یا وضعیت نامعتبر/پس‌گرفته‌شده
۵۸	**جمع**

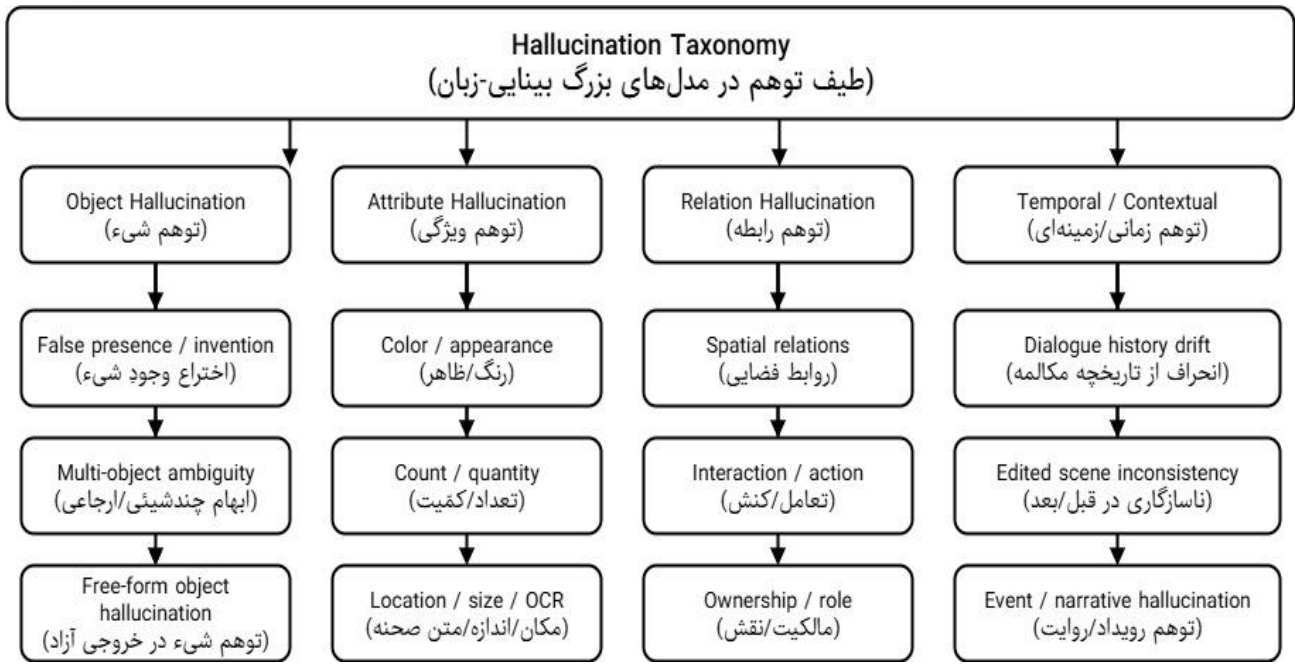
۳- طبقه‌بندی آسیب‌پذیری‌ها

۳-۱- طیف توهم

توهم در مدل‌های بزرگ بینایی-زبان را می‌توان به‌عنوان ناهمخوانی قابل‌تشخیص بین خروجی زبانی و شواهد بصری/زمینه‌ای ورودی تعریف کرد که در عمل از «اختراع یک شیء» تا «ساختن یک روایت کامل» امتداد دارد [2].

اگرچه ادبیات اولیه عمدتاً توهم را در سطح شیء، ویژگی و رابطه دسته‌بندی می‌کرد، شواهد جدید نشان می‌دهد که انواع پیچیده‌تری مانند توهم رویداد/روایت نیز به‌طور معنادار رخ می‌دهد و ارزیابی‌های صرفاً شیء-محور برای پوشش آن کافی نیست [7].

از منظر «نقشه شکست»، این طبقه‌بندی سه پرسش را برای هر نوع توهم پاسخ می‌دهد: (۱) چه چیزی خراب می‌شود (واحد خطا چیست)، (۲) چرا خراب می‌شود (محرک‌های رایج و شرایط تشدیدکننده)، و (۳) چگونه اندازه‌گیری می‌شود (پروتکل‌ها و متریک‌های متداول) [19].



شکل ۲. درخت طبقه‌بندی توهم در مدل‌های بزرگ بینایی-زبان

به همین دلیل، در این بخش یک درخت طبقه‌بندی در شکل ۲ پیشنهاد می‌شود که توهم را از سطح موجودیت (Object) تا سطح ویژگی (Attribute)، رابطه (Relation) و نهایتاً زمینه/زمان (Temporal/Contextual) سازمان‌دهی می‌کند [7].

۳-۱-۱- توهم شیء

توهم شیء زمانی رخ می‌دهد که مدل، وجود یک شیء/موجودیت را در تصویر گزارش می‌کند در حالی‌که آن شیء در تصویر حاضر نیست یا مدل درباره حضور/عدم حضور آن دچار اختراع می‌شود. این نوع توهم معمولاً در پاسخ‌های توصیفی و پرسش‌وپاسخ تصویری به شکل «افزودن اشیای رایج» یا «تکمیل‌کردن صحنه با پیش‌فرض‌های زبانی» مشاهده می‌شود [2].

مطالعه‌ی لی و همکاران نشان می‌دهد که الگوی دستور و هم‌رخدادی‌های رایج در متن‌های آموزشی می‌تواند احتمال توهم شیء را افزایش دهد، به‌ویژه وقتی اشیای پرتکرار در دستور یا اشیای هم‌رخداد با صحنه فعال شوند [2].

در سناریوهای چندشیئی، فشار شناختی مدل برای هم‌زمان «تشخیص چند موجودیت» می‌تواند خطا را تشدید کند و مدل به میانبرهای آماری یا هم‌بستگی‌های کاذب متکی شود [10].

حتی طول پاسخ نیز می‌تواند یک عامل مداخله‌گر باشد، به‌طوری‌که طولانی‌تر شدن توصیف‌ها با افزایش درجه توهم شیء هم‌بستگی نشان می‌دهد و اگر کنترل نشود می‌تواند ارزیابی‌ها را ناعادلانه کند [29].

در ارزیابی‌های بسته/کنترلی، POPE یک پروتکل پرسش‌گری مبتنی بر polling ارائه می‌کند که با پرسش‌های دودوییِ متوازن

درباره حضور اشیا، سنجش پایدارتر توهم شیء را ممکن می‌سازد [2].

برای کاهش ابهام ارجاعی و تمرکز بر «موجودیت‌های متعدد»، ROPE از پرامپت‌های ارجاع بصری استفاده می‌کند و نشان می‌دهد توهم در تنظیم چندشیئی به‌طور نظام‌مند شدیدتر می‌شود [10].

برای کنترل اثر طول خروجی بر توهم، LeHaCE درجه توهم را در طول‌های یکسان نرمال‌سازی می‌کند و علاوه بر مقدار توهم، «حساسیت توهم نسبت به طول توصیف» را نیز به‌عنوان شاخص مکمل پیشنهاد می‌دهد [20].

در مقابل، THRONE نشان می‌دهد بسیاری از سنجه‌های رایج بیشتر توهم‌های «نوع II» (وابسته به قالب‌های کنترل شده مانند چندگزینه‌ای) را می‌سنجند، در حالی‌که توهم‌های «نوع I» در تولید آزاد (free-form) ممکن است الگوی متفاوتی داشته باشد و حتی با نوع II هم‌بستگی منفی نشان دهد [۱۹].

۳-۱-۲- توهم ویژگی: رنگ/تعداد/مکان/...

توهم ویژگی زمانی رخ می‌دهد که مدل یک ویژگی قابل مشاهده یا قابل استنتاج از تصویر (مانند رنگ، تعداد، اندازه، موقعیت مکانی، جنس، یا متنِ روی اشیا) را نادرست گزارش کند یا ویژگی‌ای را «بسازد» که در تصویر وجود ندارد [7].

این نوع توهم از نظر ریسک کاربردی مهم است، زیرا ممکن است مدل «شیء درست» را تشخیص دهد اما ویژگی‌های آن را غلط بیان کند و همین برای تصمیم‌گیری‌های حساس کافی است تا خروجی را غیرقابل اتکا کند[1].

توهم ویژگی می‌تواند زمانی تشدید شود که پرسش یا زمینه متنی، مدل را به سمت یک ویژگی خاص هدایت کند و مدل به جای اتکا به شواهد بصری، از پیشین‌های زبانی یا الگوهای هم‌بستگی استفاده کند [1].

سنجه‌های مبتنی بر دستکاری صحنه، نشان می‌دهند حتی وقتی تغییرات بصری واقعی رخ داده است، مدل ممکن است در تشخیص «تغییر ویژگی» ناکام بماند و پاسخ‌های نامقاوم ارائه دهد که به‌عنوان توهم ویژگی قابل تفسیر است [21].

در چارچوب‌های ریزدانه، Hal-Eval صراحتاً توهم ویژگی را به‌عنوان یک دسته مستقل در کنار توهم شیء و رابطه صورت‌بندی می‌کند و داده/پروتکل ارزیابی را برای پوشش این طیف گسترش می‌دهد [7].

در ارزیابی‌های «استدلال مبتنی بر زمینه تصویر»، HallusionBench با طراحی جفت‌پرسش‌های کنترل شده، امکان تحلیل تمایلات پاسخ‌دهی و شکست‌های ظریف از جمله خطاهای ویژگی‌محور را فراهم می‌کند [1].

در رویکردهای مبتنی بر تغییر صحنه، BEAF با ایجاد تصاویر قبل/بعد و تعریف متریک‌های change-aware، نشان می‌دهد برخی پاسخ‌هایی که در سنجه‌های متنی-محور «غیرتوهمی» تلقی می‌شوند، در برابر تغییرات ویژگی عملاً توهم محسوب می‌گردند [21].

۳-۱-۳- توهم رابطه: روابط و تعاملات

توهم رابطه زمانی رخ می‌دهد که مدل رابطه میان موجودیت‌ها را نادرست گزارش کند، از جمله روابط فضایی (چپ/راست/بالا/پایین)، روابط کارکردی/تعامل (گرفتن، خوردن، نگاه‌کردن)، یا روابط مالکیت/وابستگی که از تصویر قابل استنتاج است [7].

این نوع توهم اغلب در پرسش‌های استدلالی شدیدتر می‌شود، زیرا مدل باید چند موجودیت را هم‌زمان ردیابی کند و سپس یک گزاره رابطه‌ای تولید کند که به شواهد دقیق وابسته است [1].

در تنظیم‌های چندشیئی، مدل ممکن است برای کاهش پیچیدگی، برخی موجودیت‌ها را نادیده بگیرد یا روابط را به‌صورت پیش‌فرض‌های کلی بازسازی کند که به توهم رابطه منجر می‌شود [1].

در تعامل‌های چندمرحله‌ای، تاریخچه متنی می‌تواند به‌عنوان «زمینه گمراه‌کننده» عمل کند و مدل را به سمت روابطی سوق دهد که با تصویر فعلی سازگار نیست [8].

در ارزیابی‌های ریزدانه، Hal-Eval توهم رابطه را در کنار سایر انواع توهم می‌سنجد و نشان می‌دهد که تمرکز صرف بر اشیا، شکست‌های رابطه‌ای را پنهان می‌کند [7].

در ارزیابی‌های مبتنی بر زمینه و کنترل، HallusionBench با ساختار جفت‌پرسش و تحلیل سازگاری منطقی، امکان مشاهده شکست‌های رابطه‌ای را به‌ویژه در پرسش‌های نیازمند استدلال فراهم می‌کند [1].

در سناریوهای مبتنی بر دیالوگ، VisDiaHalBench با تکیه بر grounding در scene graph و تاریخچه چندنوبتی، امکان تفکیک دقیق‌تر خطاهای رابطه‌ای ناشی از «تاریخچه» را فراهم می‌کند [8].

۳-۱-۴- توهم زمانی/زمینه‌ای

توهم زمانی/زمینه‌ای به شکست‌هایی اشاره دارد که در آن مدل، گزاره‌هایی درباره توالی رخدادها، تغییرات در زمان، یا وضعیت‌های وابسته به زمینه (مانند تاریخچه مکالمه یا نسخه‌های ویرایش‌شده صحنه) تولید می‌کند که با شواهد زمانی/زمینه‌ای ورودی ناسازگار است. در سطح تصویر-محور، این مسئله می‌تواند به شکل «پایبندی به یک تاریخچه گمراه‌کننده» یا ناتوانی در تشخیص تغییرات قبل/بعد ظاهر شود که به‌طور عملی به پاسخ‌های نادرست و بعضاً با اعتمادبه‌نفس بالا منجر می‌شود.

در سطح پیچیده‌تر، Hal-Eval نشان می‌دهد که توهم می‌تواند به «ساختن یک رویداد یا روایت پیرامون یک موجودیت خیالی» گسترش یابد که با دسته‌بندی‌های کلاسیک شیء/ویژگی/رابطه به‌طور کامل پوشش داده نمی‌شود [7].

چگونه اندازه‌گیری می‌شود؟ (نمونه‌های نماینده در ادبیات جدید)

در حوزه ویدئو، VidHal یک سنجه برای توهم‌های مبتنی بر جنبه‌های زمانی ارائه می‌کند و با طراحی وظیفه «رتبه‌بندی کپشن‌ها بر حسب میزان توهم»، ارزیابی ریزدانه‌تری از ناسازگاری‌های زمانی فراهم می‌کند. VidHalluc نیز با ساخت ویدئوهای جفت‌شده و تمرکز بر سه بعد «کنش، توالی زمانی و گذار صحنه»، نشان می‌دهد مدل‌ها در تمایز ویدئوهای شبیه از نظر معنا اما متفاوت از نظر جزئیات زمانی، مستعد توهم هستند [14].

در سطح تصویر اما با «پرامپت‌های زمینه‌ای»، Hallucinogen از پرامپت‌های استدلال زمینه‌ای به‌عنوان حمله‌های توهم استفاده می‌کند تا خطاهای حضور/عدم حضور اشیا تحت زمینه‌های پیچیده را آشکار کند [3].

این زیرگونه‌ها نشان می‌دهند که حرکت از تصویر منفرد به دیالوگ چندنوبتی یا ویدئو، سطح شکست را از «موجودیت‌های ثابت» به «حالت‌های وابسته به زمان و زمینه» گسترش می‌دهد و نیازمند سنجه‌ها و متریک‌های خاصِ زمان/زمینه است [۸], [۲۱], [۲۴].

در جدول ۶، فراوانی انواع توهم نه بر اساس «نرخ خطا» بلکه بر اساس «پوشش ادبیات» گزارش می‌شود؛ یعنی واحد شمارش، مطالعه است (N=18) و هر مطالعه بر اساس نوع(های) توهمی که به‌صورت صریح تعریف/هدف‌گذاری کرده یا به‌صورت تجربی با سنجه‌های متناظر ارزیابی شده، در یک یا چند رده کدگذاری شده است.

چهار رده عملیاتی برای کدگذاری شامل:

۱. توهم شیء (ادعای وجود موجودیت/شیء غایب)،
۲. توهم ویژگی (رنگ/تعداد/متنِ صحنه/OCR/مقدار عددی)،
۳. توهم رابطه (روابط مکانی/تعاملات/وابستگی‌های بین اشیاء)، و
۴. توهم زمینه‌ای/زمانی (ساخت روایت/رویداد، وابستگی به تاریخچه چندمرحله‌ای، یا خطاهای ناشی از تعارض زمینه و ورودی بصری) است. رده «رویداد» در Hal-Eval ذیل رده زمینه‌ای/زمانی لحاظ شده و مطالعات چندمرحله‌ای مانند VisDiaHalBench نیز در همین رده کدگذاری شده‌اند.

برای مطالعاتی که روش کاهش[1] توهم ارائه می‌دهند، مبنای کدگذاری عمدتاً سنجه(های) گزارش‌شده در ارزیابی است؛ به‌عنوان مثال، AMBER صراحتاً شامل توهم‌های شیء/ویژگی/رابطه است، بنابراین روش‌هایی که روی AMBER گزارش می‌دهند در هر سه رده علامت‌گذاری می‌شوند.

جدول ۶. فراوانی انواع توهم در مطالعات مرورشده با تمرکز توهم (N = ۱۸)

نوع توهم	دامنه عملیاتی (معیار کدگذاری)	تعداد مطالعات (N=۱۸)	آثار نمایندگی (نمونه)
توهم شیء	ادعای وجود/هویت شیء ناموجود یا جایگزینی نادرست شیء	15	POPE; THRONE; ROPE; LeHaCE; Hallucinogen; VisDiaHalBench; BEAF; also, mitigation methods such as OPA-DPO, V-DPO, TPR, HIO, Rep. Editing
توهم ویژگی	خطا در ویژگی‌ها (رنگ/شمارش/کمیت/متن صحنه/صفات)	7	Hal-Eval; HallusionBench; TextHalu-Bench; Rawte و همکاران; also, LLaVA-RLHF, OPA-DPO (AMBER), V-DPO
توهم رابطه	خطا در روابط فضایی/تعاملات/وابستگی‌های بین اشیاء	5	Hal-Eval; LLaVA-RLHF (MMHal-Bench: Relation); OPA-DPO (AMBER: relation); V-DPO (relationships); DSCR (spatial grounding)
توهم زمانی/زمینه‌ای	خطاهای مرتبط با تاریخچه، ماهیت چندمرحله‌ای، یا سناریوهای قبل/بعد	6	VisDiaHalBench (multi-turn history); BEAF (before/after changes); Hal-Eval (event); Hallucinogen (contextual reasoning prompts); HallusionBench; Rawte و همکاران
شمارش‌ها چندبرچسبی هستند؛ یک مطالعه ممکن است همزمان چند نوع توهم را پوشش دهد، بنابراین جمع ردیف‌ها برابر N نیست. همچنین، این جدول روی ۱۸ مطالعه‌ی «توهم‌محور» (سنجه/ارزیابی + کاهش توهم) ساخته شده است. آثار «ایمنی/گریز امنیتی » و «تحلیل ریشه‌ای» در این جدول لحاظ نشده‌اند و در بخش‌های مربوطه پوشش داده می‌شوند.			

[1] mitigation

۳-۲- ایمنی و پایداری در برابر حمله

این زیربخش آسیب‌پذیری‌های ایمنی در LVLM/MLLMها را به‌عنوان یک «مسیر شکست» صورت‌بندی می‌کند، یعنی از تعریف مدل تهدید و سطح دسترسی مهاجم تا الگوهای رایج گریز امنیتی بصری و شاخص‌های تجربی موفقیت حمله [22], [23]. در ادبیات اخیر، شواهد نشان می‌دهد که حتی مدل‌هایی که در سطح زبانی همترازسازی ایمنی شده‌اند، می‌توانند با ورودی‌های چندوجهی به تولید خروجی ناایمن سوق داده شوند[14], [15].

۳-۲-۱- مدل تهدید

در سناریوی جعبه‌سیاه مهاجم تنها به ورودی/خروجی مدل دسترسی دارد و بدون مشاهده پارامترها یا گرادیان‌ها، حمله را با طراحی پرامپت و ورودی بصری پیش می‌برد [5], [23].

بخش قابل‌توجهی از حملات عملی گزارش‌شده علیه مدل‌های بزرگ بینایی-زبان[1] در همین چارچوب جعبه‌سیاه تعریف شده‌اند، زیرا بسیاری از سامانه‌های تجاری امکان دسترسی داخلی را فراهم نمی‌کنند[32] ,[19].

در سناریوی gray-box مهاجم معمولاً به برخی اطلاعات محدود (مانند معماری کلی، نوع رمزگذار، یا رفتارهای قابل پیش‌بینی فیلترها) دسترسی دارد و از آن برای افزایش پایداری حمله یا انتقال‌پذیری بین مدل‌ها استفاده می‌کند [23].

ساختار ورودی (تک‌تصویر vs چندتصویر/ویدئو vs چندنوبتی).

بسیاری از سنجه‌های ایمنی چندوجهی حمله را در قالب یک تصویر + یک متن مدل می‌کنند، زیرا این ساده‌ترین شکل برای سنجش نفوذپذیری و محاسبه نرخ موفقیت حمله است[15].

با این حال، ادبیات جدید نشان می‌دهد که چندتصویر/ویدئو می‌تواند سطح حمله را گسترش دهد و مسیرهای جدیدی برای دورزدن همترازی ایجاد کند، به‌ویژه وقتی مدل برای ادراک زمانی و تمرکز بصری متکی به سازوکارهای توجه است[24].

همچنین حملات چندنوبتی با استفاده از تاریخچه مکالمه می‌توانند «رانش ایمنی» ایجاد کنند و در برخی چارچوب‌ها به‌عنوان یک سطح تهدید مستقل گزارش شده‌اند[15].

هدف حمله و معیارهای موفقیت.

در اکثریت مطالعات، هدف حمله «کاهش نرخ امتناع و افزایش تولید محتوای ناایمن» تعریف می‌شود و معیار رایج گزارش‌دهی، نرخ موفقیت حمله است[15] ,[14].

برخی کارها برای لحاظ‌کردن چند الگوی حمله/چند قالب پرامپت، معیارهای تجمیعی مانند نرخ موفقیت حمله گروهی[2] را نیز گزارش می‌کنند تا اثر «چندبار تلاش» و «تنوع قالب‌ها» در موفقیت نفوذ کمی‌سازی شود[6].

۳-۲-۲- گریز امنیتی بصری (حروف‌نگاری /نویز/تضاد بین حالت‌ها/حواس‌پرتی)[3]

- حملات کانال حروف‌نگاری /OCR[4]

یک خط حمله مؤثر، تبدیل محتوای ممنوع به نمایش تصویری-نوشتاری است تا سیاست‌های ایمنی متن‌محور دور زده شود و مدل از کانال بینایی آن را بازخوانی کند [5], [23].

روش فیگ‌استپ - که در آن محتوای مخرب متنی به‌صورت تصویر حروف‌نگاری‌شده رندر شده و به‌جای ورودی متنی به مدل داده می‌شود - نشان می‌دهد چنین حمله‌ای می‌تواند در محیط جعبه‌سیاه نرخ موفقیت بالایی داشته باشد و به‌طور متوسط نرخ موفقیت حمله قابل‌توجهی را روی چند مدل بزرگ بینایی-زبان متن‌باز گزارش می‌کند [5], [23].

تصویر مرتبط با جستجو / حامل با ظاهری بی‌خطر[5]

دسته‌ای دیگر از حملات، استفاده از تصاویر ظاهراً مرتبط و بی‌خطر است که وقتی با پرسش مخرب جفت می‌شوند، می‌توانند سازوکارهای ایمنی را تضعیف و خروجی ناایمن را فعال کنند[15].

معیار ایمنی چندوجهی این الگو را به‌صورت نظام‌مند ارزیابی می‌کند و نشان می‌دهد مدل‌های چندوجهی در برابر این نوع دستکاری‌ها حتی با وجود همترازسازی زبانی، شکننده باقی می‌مانند[15].

- ساختن تصویر برای پنهان/تقویت نیت مخرب[6]

چارچوب هیدیز[7]-که از راهبرد دومرحله‌ای پنهان‌سازی نیت مخرب

[1] Large Vision-Language Model (LVLM)
[2] Ensemble ASR (EASR)
[3] Visual Jailbreaking (Typographic/Noise/Cross-modal conflict/Distraction)
[4] Typographic/OCR channel attacks
[5] Amplification & concealment
[6] HADES (Hiding And Deliberately Evoking Subtly)

در تصویر و سپس تقویت آن در مرحله پاسخ‌دهی بهره می‌برد - نشان می‌دهد می‌توان هم‌ترازی بی‌ضرری را دور زد و نرخ موفقیت حمله بالایی را روی چند مدل گزارش کرد. این نتیجه از منظر مدل تهدید اهمیت دارد، زیرا حمله می‌تواند بدون تغییر معماری یا دسترسی داخلی، صرفاً با مهندسی ورودی چندوجهی اجرا شود[15].

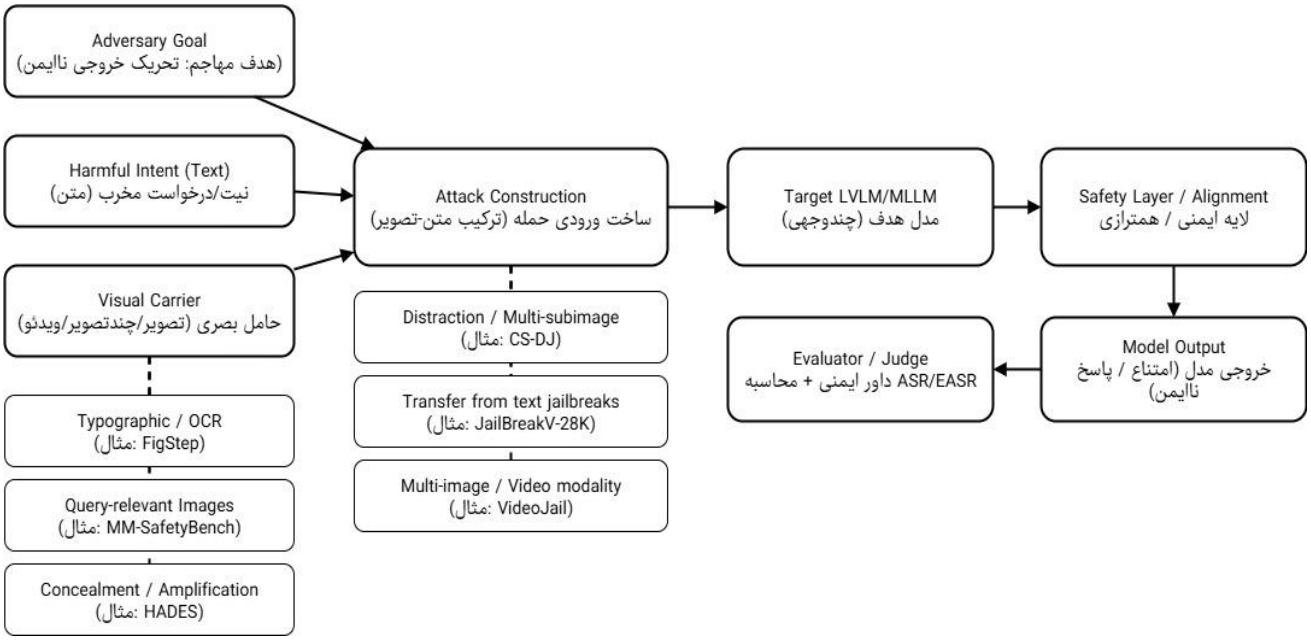
حملات تعاملی حواس‌پرتی / چند زیرتصویری / بین‌حالتی[1]

یک مسیر پررنگ در کارهای جدید، بهره‌گیری از برهم‌کنش‌های پیچیده بین زیرتصویرها و توجه بصری برای کاهش توان مدل در تشخیص نیت مخرب و افزایش نرخ عبور از فیلترهای ایمنی است[6].

CS-DJ این ایده را با «حواس‌پرتی ساختاری» و «حواس‌پرتی تقویتی بصری» عملیاتی می‌کند و نشان می‌دهد پیچیدگی و سازمان‌دهی زیرتصویرها می‌تواند با موفقیت حمله هم‌بستگی مستقیم داشته باشد[6].

- انتقال حملات از گریز امنیتی های متنی به چندوجهی

گریز امنیتی وی-۲۸هزار نشان می‌دهد بخش معناداری از تکنیک‌های گریز امنیتی متنی می‌توانند به مدل‌های بزرگ زبانی چندوجهی[2] منتقل شوند و این انتقال‌پذیری یک آسیب‌پذیری «ترکیبی» ایجاد می‌کند که هم از متن و هم از تصویر تغذیه می‌شود[5].



شکل ۳. خط لوله گریز امنیتی بصری

شکل ۳، خط لوله مفهومی گریز امنیتی بصری نشان می‌دهد که چگونه نیت مخرب می‌تواند از طریق کانال بینایی (برای مثال به‌صورت محتوای حروف‌نگاری‌شده، تصاویر ظاهراً مرتبط، یا حواس‌پرتی چندزیرتصویر) در ورودی چندوجهی ادغام شود و با عبور از لایه‌های همترازی، به خروجی ناایمن منجر گردد. این چارچوب همچنین جایگاه ارزیابی تجربی را با محاسبه ASR/EASR پس از داوری ایمنی مشخص می‌کند و امکان مقایسه نظام‌مند خانواده‌های حمله را فراهم می‌سازد.

۳-۲-۳- جمع‌بندی تجربی

الگوی اول: حملات ساده اما پایدار در جعبه‌سیاه می‌توانند نرخ موفقیت بالا داشته باشند.

FigStep به‌عنوان یک حمله جعبه‌سیاه مبتنی بر حروف‌نگاری، نشان می‌دهد انتقال محتوا به کانال بینایی می‌تواند همترازی متن‌محور را دور بزند و نرخ موفقیت حمله متوسط بالایی را روی چند مدل متن‌باز گزارش می‌کند [5], [23].

به‌طور مشابه، هیدیز نشان می‌دهد «تصویر» می‌تواند به‌عنوان یک ابزار تقویت/پوشانندگی برای نیت مخرب عمل کند و نرخ موفقیت حمله متوسط بسیار بالا برای برخی مدل‌ها گزارش شده است.

الگوی دوم: تصاویر ظاهراً مرتبط می‌توانند شکست همترازی را حتی در مدل‌های همترازشده فعال کنند.

[1] Distraction / multi-subimage / cross-modal interaction attacks

[2] Multimodal Large Language Model(MLLM)

معیار ایمنی چندوجهی نشان می‌دهد که جفت‌کردن پرسش مخرب با تصاویر query-relevant می‌تواند منجر به نقض سیاست شود، حتی وقتی مدل پایه زبانی از پیش تحت همترازی قرار گرفته است[15].

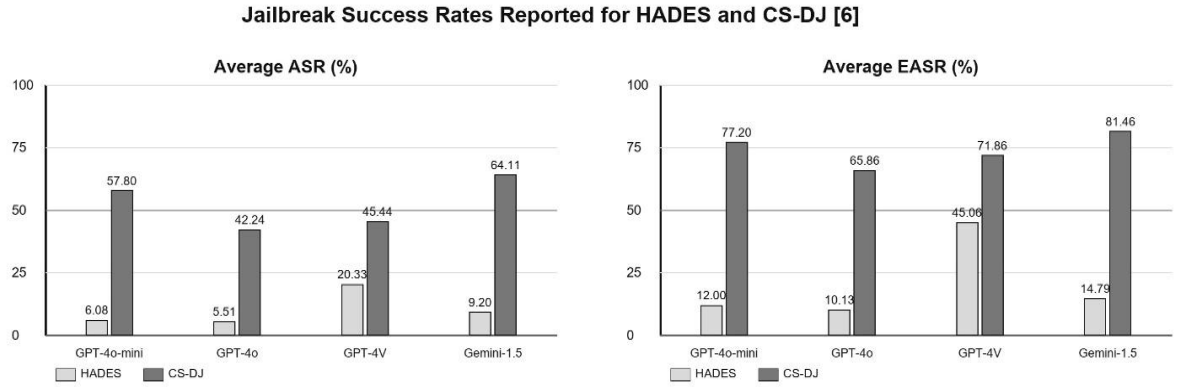
الگوی سوم: «حواس‌پرتی» می‌تواند نرخ نفوذ را در مدل‌های تجاری نیز بالا ببرد و تفاوت‌های بین مدل‌ها را آشکار کند.

CS-DJ در مقایسه با هیدیز، برای چهار مدل تجاری (GPT-4o-mini، GPT-4o، GPT-4V، Gemini-1.5-Flash) افزایش معنی‌دار نرخ موفقیت حمله و EASR را گزارش می‌کند و میانگین نرخ موفقیت حمله را در کل آزمایش‌ها ۵۲٫۴۰٪ و میانگین EASR را ۷۴٫۱۰٪ بیان می‌کند[6].

نتایج تفکیکی CS-DJ نشان می‌دهد که آسیب‌پذیری ایمنی می‌تواند بین حوزه‌های آسیب (مانند مالی/حریم خصوصی/خشونت) و نیز بین مدل‌ها متفاوت باشد و این ناهمگونی برای طراحی دفاع‌های تعمیم‌پذیر حیاتی است[6].

الگوی چهارم: کانال‌های ورودی غنی‌تر (ویدئو) می‌توانند سطح حمله را تشدید کنند.

VideoJail نشان می‌دهد افزوده‌شدن ماهیتی ویدئو می‌تواند مسیرهای جدیدی برای شکست بی‌ضرری ایجاد کند و برای برخی مدل‌ها نرخ موفقیت حمله بسیار بالا گزارش می‌شود[24].



شکل ۴. نرخ موفقیت گریز امنیتی CS-DJ براساس نتایج گزارش‌شده در[6]

شکل ۴ مقایسه نرخ موفقیت حمله و نرخ موفقیت تجمیعی (EASR) برای دو روش هیدیز و CS-DJ روی چهار مدل تجاری نشان می‌دهد که روش‌های مبتنی بر حواس‌پرتی چندزیرتصویر می‌توانند شکاف ایمنی قابل‌توجهی ایجاد کنند. ارقام شکل مستقیماً از نتایج میانگین گزارش‌شده در جدول نتایج CS-DJ استخراج شده‌اند و برای تحلیل تفاوت بین مدل‌ها و نیز تفاوت بین معیار نرخ موفقیت حمله و EASR استفاده می‌شوند[6].

۳-۳- تحلیل ریشه‌ای (Root Cause Analysis)

در حالی که بخش‌های ۳-۱ و ۳-۲ مشخص می‌کنند «چه چیزی» در خروجی مدل‌ها دچار شکست می‌شود، این زیربخش به «چرایی» این شکست‌ها می‌پردازد و علل را در سه لایه‌ی بازنمایی[1]، داده[2]، و معماری/استنتاج[3] سازمان‌دهی می‌کند[8] ,[2].

این تفکیک به‌صورت مستقیم با نقشه خرابی[4] هم‌راستا است، زیرا برای هر علت ریشه‌ای می‌توان (الف) نشانه‌های شکست، (ب) شاخص‌های اندازه‌گیری، و (ج) نقاط مداخله[5] را مشخص کرد [25].

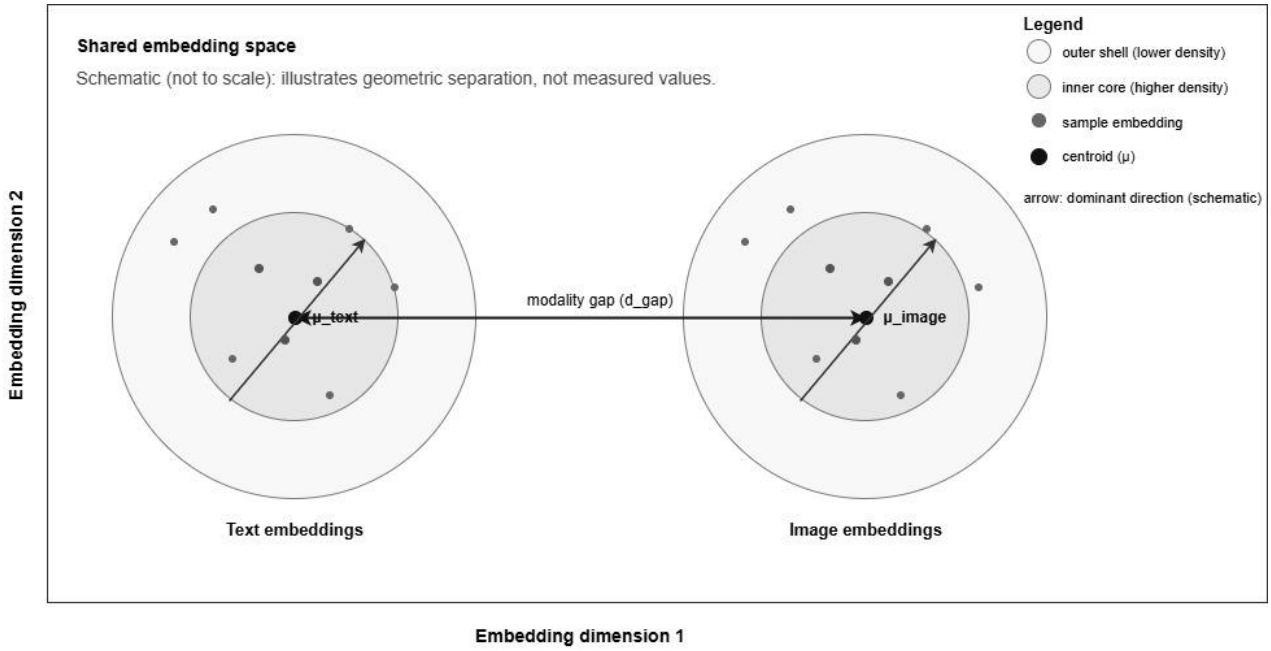
۳-۳-۱- شکاف ماهیتی (Modality Gap)

شکاف ماهیتی به پدیده‌ای اشاره دارد که در آن بازنمایی‌های تصویری و متنی، علی‌رغم آموزش چندوجهی، در فضای نهفته به‌طور کامل هم‌پوشان نمی‌شوند و عملاً در دو ناحیه نسبتاً جدا قرار می‌گیرند. Liang و همکاران نشان می‌دهند این شکاف یک ویژگی هندسی در مدل‌های کنتراستیو مانند CLIP است و می‌تواند ناشی از ترکیب مقداردهی اولیه و دینامیک بهینه‌سازی یادگیری کنتراستیو باشد[8].

از منظر قابلیت اعتماد، پیامد عملی این شکاف آن است که در

[1] representation
[2] data
[3] architecture & inference
[4] Failure Map
[5] handles

شرایط ابهام یا سیگنال بصری ضعیف، مدل مستعد اتکا به پیشین‌های زبانی و الگوهای آماری متن می‌شود و این اتکا می‌تواند به انواع توهم (به‌ویژه در سطح شیء/ویژگی) منتهی شود. Schrodi و همکاران این تحلیل را با نشان‌دادن پیوند مشترک بین شکاف ماهیتی و «سوگیری شیء» تقویت می‌کنند و استدلال می‌کنند عامل محرک هر دو، «عدم‌تعادل اطلاعاتی» بین تصاویر و کپشن‌ها در داده/فرآیند آموزش است. بر مبنای این نتیجه، شکاف ماهیتی صرفاً یک پدیده‌ی زیباشناختی در embedding نیست، بلکه یک سیگنال تشخیصی برای شرایطی است که احتمال برتری یافتن کانال زبانی بر کانال بصری افزایش می‌یابد[9].



شکل ۵. نمایش شماتیک شکاف ماهیتی در فضای تعبیه مشترک

همان‌گونه که در شکل ۵ نشان داده شده است، تعبیه‌های متن و تصویر به‌صورت دو خوشه‌ی نیمه‌جدا ظاهر می‌شوند؛ فاصله‌ی میان مراکز خوشه‌ها (μ_text و μ_image) به‌عنوان d_gap نمایش داده شده است. هسته‌های داخلی نشان‌دهنده‌ی نواحی پرتراکم‌تر و پیکان‌ها بیانگر جهت‌های غالب زیرفضا هستند که می‌توانند جدایی میان ماهیتی‌ها را حفظ کنند.

۳-۳-۲- سوگیری داده، هم‌رخدادی، و نویز کپشن

بخش قابل‌توجهی از توهم‌ها را می‌توان به سوگیری‌های آماری داده نسبت داد، به‌طوری‌که مدل به‌جای استنتاج مبتنی بر شواهد بصری، از هم‌رخدادی‌های پرتکرار و میانبرهای هم‌بستگی استفاده می‌کند. در مطالعه POPE نشان داده می‌شود اشیایی که (الف) در دستورهای بصری پرتکرارترند یا (ب) با اشیای حاضر در تصویر هم‌رخدادی رایج دارند، به‌طور معنادار بیشتر «توهم» می‌شوند[2].

این پدیده در ادبیات کپشن‌گذاری نیز با عنوان «سوگیری شیء/توهم شیء ناشی از هم‌رخدادی» بررسی شده و روش‌هایی مانند نرمال‌سازی آمار هم‌رخدادی برای کاهش این توهم‌ها پیشنهاد شده‌اند. همچنین نتایج ICD نشان می‌دهد «اغتشاش در دستور» می‌تواند عدم‌قطعیت همترازی چندوجهی را افزایش دهد و در نتیجه مدل آسان‌تر به مفاهیم هم‌رخدادِ سوگیرانه از پیش‌آموزش متکی شود[20].

در کنار سوگیری هم‌رخدادی، کیفیت جفت‌های تصویر–متن نیز عامل کلیدی است، زیرا داده‌های وب‌مقیاس می‌توانند دارای «نویز تطابق/Noisy Correspondence» باشند و این نویز مستقیماً یادگیری همترازی بین ماهیتی‌ها را تضعیف می‌کند. کارهای جدیدتر روی «یادگیری/حذف اثر تطابق نویزی» نشان می‌دهند این نویز می‌تواند دانش نادرست یا هم‌بستگی‌های کاذب را در بازنمایی‌های چندوجهی تثبیت کند و حذف/Unlearning آن می‌تواند پایداری مدل را بهبود دهد[26].

از منظر عملی، «کپشن‌های نویزی» می‌توانند علاوه بر کاهش دقت retrieval، موجب تزریق سیگنال‌های غلط در آموزش کنتراستیو شوند و مسئله false positives/false negatives را در انتخاب

نمونه‌ها تشدید کنند. همچنین در سناریوهای انتخاب نمونه پاک (clean sample selection)، اگر مدل از ابتدا شباهت‌ها را با خطا پیش‌بینی کند، چرخه‌ای از خطاهای خود-تقویت‌شونده شکل می‌گیرد که می‌تواند همترازی را به سمت روابط نادرست سوق دهد[1]

چگونه اندازه‌گیری می‌شود؟

برای سوگیری هم‌رخدادی، سنجه‌هایی مانند co-occurrence score (بین اشیای گزارش‌شده و اشیای صحنه) و تحلیل توزیع خطاها نسبت به فراوانی اشیا/هم‌رخدادی‌ها گزارش می‌شود[2].

برای نویز تطابق، ارزیابی معمولاً شامل (الف) سنجش میزان جفت‌های غلط یا ضعیف، و (ب) اثر آن بر بازیابی/همترازی و پایداری در وظایف پایین‌دستی است[26].

۳-۳-۳- گلوگاه‌های معماری و سوگیری‌های استنتاجی[1]

در بسیاری از مدل‌های بزرگ بینایی-زبان، معماری به‌صورت «رمزگذار بینایی + پل/پروژکتور + LLM خودرگرسیو» پیاده‌سازی می‌شود و همین اتصال می‌تواند یک گلوگاه اطلاعاتی ایجاد کند که در آن جزئیات بصری به اندازه کافی به فضای زبانی تزریق نمی‌شوند. این گلوگاه در وظایف ادراکِ دقیق (مانند نمودارها و عناصر متنی/عددی) پررنگ‌تر می‌شود و به‌عنوان « perception bottleneck» گزارش شده است که ریشه آن می‌تواند در محدودیت‌های رمزگذار بینایی و مسیر انتقال ویژگی‌ها باشد [15, 33].

از سوی دیگر، حتی با ورودی بصری مناسب، فرآیند تولید خودرگرسیو می‌تواند باعث «غلبه‌ی توالی‌های زبانی محتمل» شود، به‌خصوص زمانی که مدل در grounding بصری دچار عدم‌قطعیت است و خروجی را با تکیه بر prior زبانی کامل می‌کند. IBD این مسئله را به‌صورت مستقیم با مفهوم «اتکای بیش‌ازحد به پیشین زبانی» صورت‌بندی می‌کند و نشان می‌دهد تغییر سیاست decoding به سمت تصویرمحور می‌تواند توهم را کاهش دهد[15].

ICD نیز نشان می‌دهد اغتشاش‌های کنترلی (در متن دستور یا ورودی) می‌توانند توهم را از طریق اختلال در ماژول‌های همجوشی چندوجهی تشدید کنند، که این امر به وجود نقاط شکننده معماری/استنتاج در مسیر fusion اشاره دارد. در سطح کلی‌تر، ادبیات مرورهای فنی درباره توهم در مدل‌های زبانی نیز تأکید می‌کند که عوامل استنتاجی (مانند راهبرد decoding و پدیده‌های اعتمادبه‌نفس/بیش‌اطمینانی) می‌توانند به تولید گزاره‌های نادرست با اطمینان بالا منتهی شوند، که در مدل‌های چندوجهی نیز به‌صورت هم‌افزا با ضعف grounding بصری ظاهر می‌شود[26].

چگونه اندازه‌گیری می‌شود؟

برای گلوگاه ادراکی/معماری، تحلیل معمولاً با آزمون‌های کنترل شده‌ی ادراک دقیق، تحلیل نقش رمزگذار بینایی، و مطالعه‌ی مسیرهای attention/همجوشی انجام می‌شود. برای سوگیری استنتاجی، مقایسه‌ی خروجی تحت decodingهای مختلف یا مداخلات inference-time (مانند decoding تصویرمحور یا contrastive decoding) به‌عنوان ابزار تشخیصی/کاهشی گزارش می‌شود [14]. جمع‌بندی علل ریشه‌ای توهم و شکست‌های ایمنی، سازوکارهای ایجاد آن‌ها، شاخص‌های تشخیصی و نقاط مداخله مرتبط در جدول ۷ ارائه شده است.

جدول ۷. ماتریس علل ریشه‌ای

علت ریشه‌ای	مکانیسم (چرا ایجاد می‌شود؟)	شکست‌های غالب (3.2 / 3.1)	شاخص‌های تشخیصی/اندازه‌گیری	نقاط مداخله رایج (Training/Inference)
Modality Gap	جدایی بازنمایی تصویر/متن و ادغام ناقص در فضای نهفته	افزایش اتکا به prior زبانی ⟵توهم شیء/ویژگی؛ زمینه‌ساز شکنندگی ایمنی	فاصله/جدایی embeddingها؛ تحلیل ابعاد غالب شکاف	همترازی بهتر، کاهش عدم‌تعادل اطلاعاتی
Co-occurrence & Reporting Bias	یادگیری میانبرهای آماری از هم‌رخدادی‌های پرتکرار	توهم شیء و خطاهای ویژگی‌محور	co-occurrence score؛ خطا نسبت به فراوانی/هم‌رخدادی	داده‌افزایی/نرمال‌سازی هم‌رخدادی؛ decodingمقاوم
Caption/Pair Noise (Noisy Correspondence)	جفت‌های تصویر–متن غلط/ضعیف در داده وب‌مقیاس	همترازی ضعیف ⟵افزایش خطاهای grounding و توهم	نرخ/اثر جفت نویزی؛ false pos/neg در کنتراستیو	فیلتر/پاکسازی؛ noise-robust training؛ unlearning
Architectural/Inference Bottlenecks	گلوگاه پل بینایی–زبان +تولید خودرگرسیو و غالب شدن prior زبانی	توهم در تولید آزاد؛ شکست ادراک دقیق؛ دورزدن ایمنی با ورودی‌های خاص	تست ادراک دقیق؛ تحلیل attention/fusion؛ مقایسه decoding	تقویت ادراک؛ decoding تصویرمحور/contrastive

[1] Architectural Bottlenecks

۴- ارزیابی اکوسیستم (بحران اندازه‌گیری)

۴-۱- معیارهای تمایز[1]

سنجه‌های تمایزی معمولاً ارزیابی توهم را به یک وظیفه‌ی تصمیم‌گیری محدود می‌کنند (مثلاً «آیا شیء X در تصویر هست یا نه؟») تا نمره‌دهی ساده، سریع و کم‌هزینه شود؛ اما همین ساده‌سازی می‌تواند دو اثر جانبی ایجاد کند: (۱) مدل را به سمت یک الگوی پاسخ نظام‌مند سوق دهد و (۲) بخشی از توهم‌هایی را که در تولید آزاد رخ می‌دهند، پنهان کند. در POPE، توهم شیء به‌صورت یک مسئله‌ی Yes/No صورت‌بندی می‌شود و مطالعه نشان می‌دهد که ویژگی‌هایی مثل «سوگیری پاسخ مثبت» و نیز «تأثیر دستورهای بصری/هم‌رخدادی داده» می‌تواند روی نرخ خطا اثر بگذارد؛ بنابراین نمره‌ی POPE علاوه بر «توهم»، بخشی از «سوگیری پاسخ‌دهی» و «ساختار پروبینگ» را هم اندازه می‌گیرد.[2]

این مسئله در کارهای بعدی با تمرکزهای خاص‌تر برجسته‌تر می‌شود: ROPE نشان می‌دهد توهم چند-شیئی و ابهام در ارزیابی (وقتی چند شیء/کاندیدا مطرح است) می‌تواند باعث ناپایداری نتیجه شود و نیاز به پروتکل‌های دقیق‌تر دارد[34]. همچنین LeHaCE نشان می‌دهد در ارزیابی توهمِ توصیف تصویر، طول پاسخ با درجه توهم همبستگی دارد؛ در نتیجه اگر کنترل نشود، مدل‌هایی که «مفصل‌تر» تولید می‌کنند ممکن است ناعادلانه بدتر ارزیابی شوند (یا برعکس، با کوتاه‌نویسی مصنوعی بهتر شوند) [20].

۴-۲- ارزیابی مولد

ارزیابی مولد تلاش می‌کند توهم را در شرایط نزدیک‌تر به استفاده‌ی واقعی بسنجد، یعنی وقتی مدل پاسخ آزاد و چندجمله‌ای تولید می‌کند؛ اما این خانواده معمولاً با دو مشکل روبه‌رو است: (۱) استخراج «ادعاها» از متن آزاد و (۲) داوری این ادعاها نسبت به تصویر. THRONE دقیقاً با همین انگیزه طراحی شده است: تمرکز روی تولید آزاد و ارائه‌ی چارچوب خودکار برای سنجش توهم در خروجی‌های free-form، با این نکته که نشان می‌دهد کاهش یک نوع توهم لزوماً به کاهش نوع دیگر منجر نمی‌شود و حتی می‌تواند ضدهمبستگی رخ دهد. نقطه‌ی حساس اینجاست که THRONE برای شناسایی توهم در پاسخ‌های آزاد از «مدل‌های زبانی عمومی» به‌عنوان داور/استخراج‌کننده استفاده می‌کند؛ بنابراین بخشی از تکرارپذیری نتیجه به پایداری داور بیرونی (نسخه/ تنظیمات/ دسترسی) و سوگیری‌های ارزیاب وابسته می‌شود[19].

در همین مسیر، Hallucinogen ارزیابی را به‌صورت «حمله ارزیابی» صورت‌بندی می‌کند و نشان می‌دهد طراحی پرامپت‌ها می‌تواند مدل را به سمت خطاهای خاص سوق دهد؛ بنابراین بخشی از خروجی ارزیابی، نه فقط «کیفیت مدل» بلکه «آسیب‌پذیری مدل نسبت به تحریک/القای توهم» را اندازه می‌گیرد[3].

همزمان، برخی کارها (مثل TrustNLP و Hal-Eval) تلاش می‌کنند تعریف‌ها و سطوح توهم را ریزدانه کنند تا داوری از حالت مبهم خارج شود؛ اما این ریزدانه‌سازی وقتی ارزش دارد که با پروتکل‌های استاندارد سنجش و گزارش‌دهی همراه شود، وگرنه صرفاً به «ناهمخوانی تعاریف بین مقالات» می‌انجامد [7].

در جدول ۸، منظور از «بسته» در ستون «قالب پاسخ»، ارزیابی‌هایی است که پاسخ در گزینه‌های محدود مانند Yes/No یا انتخاب چندگزینه‌ای محصور است؛ «باز» به تولید آزاد متن یا پاسخ توصیفی بدون گزینه‌های از پیش تعیین‌شده اشاره دارد؛ و «ترکیبی» پروتکل‌هایی را نشان می‌دهد که هم پاسخ بسته/کنترل شده و هم پاسخ تولیدی یا داوری آزاد را در کنار هم به‌کار می‌گیرند.

۴-۳- معیارهای ایمنی

در حوزه ایمنی، مسئله‌ی سنجش حتی پیچیده‌تر می‌شود چون خروجی مطلوب فقط «درست بودن» نیست، بلکه «پرهیز از پاسخ مضر» و «پایداری در برابر دستکاری چندوجهی» است[14] ,[5].

معیار ایمنی چندوجهی نشان می‌دهد که حتی وقتی مدل زبانی پایه همتراز شده است، اضافه‌شدن تصویر می‌تواند مسیرهای جدیدی برای شکستن همترازی ایجاد کند؛ در این سنجه، تصاویر مرتبط با کوئری می‌توانند مدل را به پاسخ‌دادن به درخواست‌های نامجاز ترغیب کنند و ارزیابی روی مجموعه سناریوهای متنوع گزارش می‌شود[14].

گریز امنیتی وی-۲۸هزار دقیقاً روی مسئله‌ی «انتقال‌پذیری گریز امنیتی های متنی به مدل‌های چندوجهی» تمرکز دارد و با ساخت یک مجموعه بزرگ از آزمون‌ها (ترکیب گریز امنیتی های متنی و ورودی‌های تصویری)، نرخ موفقیت حمله را به‌عنوان شاخص مرکزی گزارش می‌کند و آسیب‌پذیری مدل‌ها را در سناریوهای مختلف نشان می‌دهد[5].

[1] Discriminative Benchmarks

جدول ۸. معیارهای توهم

Benchmark / Suite	نوع ارزیابی	دامنه توهم	قالب پاسخ	داوری/اسکورینگ	مسئله اندازه‌گیری که برجسته می‌کند	مرجع
POPE	تمایزی (Yes/No probing)	شیء	بسته	دقت/نرخ خطا در پرسش‌های باینری	Yes-biasو حساسیت به طراحی پروبینگ/دستور	[2]
THRONE	مولد (free-form)	عمدتاً شیء(Type I در تولید آزاد)	باز	استخراج/داوری با LMهای عمومی	وابستگی به داور بیرونی و تفاوت تولید آزاد با ارزیابی بسته	[19]
LeHaCE	مولد/تحلیلی	شیء	باز	چارچوب ارزیابی پایدارتر + تحلیل طول پاسخ	همبستگی طول پاسخ و توهم (ریسک ناعادلانه‌شدن مقایسه)	[20]
ROPE	پروتکل/ارزیابی	چند-شیئی	بسته/کنترل شده	کاهش ابهام در ارزیابی چندشیئی	ابهام چند-شیئی و شکنندگی معیارهای ساده	[40]
Hal-Eval	فریم‌ورک ریزدانه	شیء/ویژگی/رابطه/رویداد	ترکیبی	سنجش چندسطحی	ناهمخوانی تعاریف و نیاز به سنجش ریزدانه	[7]
VisDiaHalBench	سنجه گفت‌وگویی	زمینه‌ای/چندمرحله‌ای	باز	سنجش خطا تحت تاریخچه	اثر تاریخچه متنی و multi-turn روی توهم	[8]
BEAF	قبل/بعد (Before-After)	زمینه‌ای +شیء/ویژگی	باز/کنترل شده	سنجش تغییرات	حساسیت به تغییرات و ادراک «تفاوت»	[21]
TextHalu-Bench	دامنه‌محور (Scene Text)	ویژگی (متن/معنا)	باز	سنجش خطا در متن صحنه	ضعف‌های domain-specific (متن/عدد)	[35]
TrustNLP (Defining/ Quantifying)	تعریف/کدگذاری	ریزدانه (چندبعدی)	ترکیبی	تعریف و کمّی‌سازی	اختلاف تعریف‌ها و مشکل قیاس‌پذیری	[11]
MMHal-Bench	سنجه توهم برای RLHF	چندبعدی (penalize hallucination)	باز	نمره‌دهی برای سناریوی کاربردی	شکاف بین معیارهای کلاسیک و سنجش کاربردمحور	[36]
AMBER / Object-Hal (به‌کاررفته)	سنجه‌های مورد استفاده در DPO	object/attribute/ relation (طبق گزارش مقاله)	ترکیبی	نرخ توهم/کاهش توهم	وابستگی نتیجه به انتخاب سنجه و تنظیمات گزارش	[38]

حمله‌محور شدن برخی ارزیابی‌ها نیز خود بخشی از اکوسیستم سنجش ایمنی است: CS-DJ نشان می‌دهد ساختاردهی تصویر به چند زیرتصویر متضاد و ایجاد حواس‌پرتی چندسطحی می‌تواند مکانیزم‌های ایمنی را دور بزند، و VoTA با زنجیره تصاویر «رقابت بین استدلال و ایمنی» را هدف قرار می‌دهد؛ بنابراین معیارهای ایمنی به‌طور مستقیم به مدل تهدید و سبک حمله وابسته‌اند [6], [25].

با گسترش ماهیتی، VideoJail نشان می‌دهد اضافه شدن ویدئو سطح حمله را بزرگ‌تر می‌کند و حتی مدل‌های تجاری/بسته هم می‌توانند در سناریوهای ویدئویی آسیب‌پذیر باشند؛ این نتیجه به‌طور مستقیم به محدودیت سنجه‌های صرفاً تصویری اشاره می‌کند[24]. در سمت دفاع/همترازی، تراز ایمنی و PSA-VLM تلاش می‌کنند ماژول‌های ایمنی را وارد مسیر بینایی کنند تا «ایمنی ماهیتی بصری» تقویت شود، و Immune و SafePTR نشان می‌دهند که بخشی از دفاع می‌تواند به زمان استنتاج/فیلترسازی لایه‌ای منتقل شود؛ اما همین تنوع روش‌ها باعث می‌شود بدون سنجه‌های مشترک و پروتکل‌های یکسان، مقایسه‌ها به‌شدت وابسته به تنظیمات ارزیابی بماند[26]. مقایسه معیارها و پروتکل‌های اصلی ارزیابی ایمنی و گریز امنیتی از نظر محور ارزیابی، مدل تهدید و نوع پوشش در جدول ۹ ارائه شده است.

جدول ۹. معیارهای ایمنی و گریز امنیتی

Benchmark / Protocol	محور اصلی	مدل تهدید	اندازه‌گیری کلیدی	پوشش ماهیتی
معیار ایمنی چندوجهی	شکستن همترازی با تصویر مرتبط با کوئری	تصویرمحور (paired text+image)	نرخ نقض ایمنی/رفتار ناامن	تصویر
گریز امنیتی وی-۲۸هزار	انتقال گریز امنیتی های متنی به MLLM	متن‌محور + تصویرمحور (ترکیبی)	Attack Success Rate (ASR)	متن+تصویر
CS-DJ	حواس‌پرتی چندزیرتصویری	تصویر چندبخشی/ترکیبی	نرخ موفقیت حمله /موفقیت گریز امنیتی	تصویر
VoTA	زنجیره تصاویر برای تحریک «فکر پرخطر»	چندتصویر/زنجیره‌ای	نرخ موفقیت حمله /رفتار ناامن	تصویر (چندمرحله‌ای)
VideoJail	آسیب‌پذیری ماهیتی ویدئو	ویدئو-محور	موفقیت گریز امنیتی /نقض دفاع	ویدئو
Immune (eval protocol)	دفاع inference-time و ارزیابی روی چند سنجه	وابسته به سنجه‌های گریز امنیتی	کاهش نرخ موفقیت حمله با حفظ utility	متن+تصویر
SafePTR (eval protocol)	دفاع training-free در سطح توکن/لایه	وابسته به سناریوهای گریز امنیتی	کاهش رفتار ناامن	متن+تصویر

جمع‌بندی شواهد این کورپس نشان می‌دهد اکوسیستم ارزیابی در دو مسیر موازی رشد کرده است: (۱) سنجه‌های توهم (عمدتاً شیء-محور یا تعاریف کلی) و (۲) سنجه‌های ایمنی/گریز امنیتی (عمدتاً سناریو-محور و تهدید-محور)، و هنوز یک سنجه «یکپارچه» که همزمان وفاداری/توهم و ایمنی/عدم‌پاسخ به محتوای مضر را تحت یک پروتکل مشترک بسنجد در این مجموعه آثار غالب نیست [4], [5], [14].

از طرف دیگر، تخصصی‌بودن سناریوها نشان می‌دهد بسیاری از سنجه‌ها پوشش محدود دارند: VisDiaHalBench و BEAF شکاف در سناریوهای چندمرحله‌ای/تغییرات قبل-بعد را برجسته می‌کنند و TextHalu-Bench نشان می‌دهد حتی اگر توهم شیء کنترل شود، خطاهای ویژگی در «متن صحنه» (که در کاربردهای واقعی رایج است) می‌تواند باقی بماند؛ بنابراین «گستره پوشش» از نظر نوع توهم و نوع داده هنوز نامتوازن است[35] ,[21] ,[8].

همچنین، در این کورپس شواهد پررنگی از سنجه‌های non-English یا سنجه‌های domain-specific فراتر از چند نمونه محدود (مانند متن صحنه) دیده نمی‌شود؛ در نتیجه، تعمیم‌پذیری نتایج به زبان‌ها و حوزه‌های تخصصی (مثلاً پزشکی/حقوقی/صنعتی) همچنان یک نقطه کور تجربی می‌ماند و باید در دستورکار پژوهش آینده قرار گیرد [30].

برای مقایسه منسجم سنجه‌های توهم و ایمنی، یک نگاشت پوشش (Coverage Map) به‌صورت Heatmap طراحی شد که در آن، هر سطر نماینده‌ی یک سنجه/پروتکل ارزیابی و هر ستون نماینده‌ی یک بُعد پوشش است. در این نگاشت، خانه‌ها به‌صورت دودویی کدگذاری شده‌اند: مقدار ۱ فقط زمانی ثبت می‌شود که آن بُعد به‌صورت صریح در تعریف سنجه، طراحی پروتکل، یا Contribution مقاله ذکر شده باشد؛ در غیر این صورت مقدار ۰ ثبت می‌شود (حتی اگر آن بُعد در بخش آزمایش‌ها به‌طور ضمنی دیده شود). این قاعده، محافظه‌کارانه است اما از نظر داوری علمی، بیشترین دفاع‌پذیری را ایجاد می‌کند.

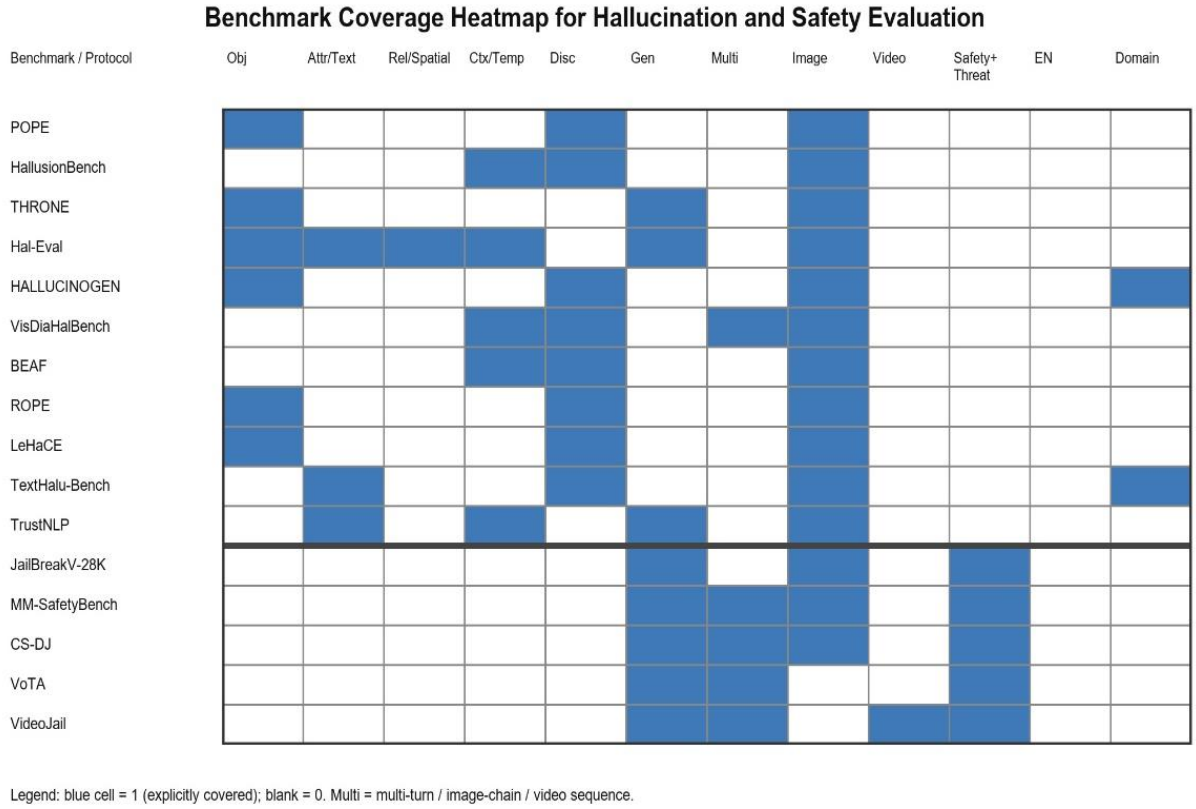
ابعاد پوشش در Heatmap شامل چهار محور توهم (توهم شیء، ویژگی/متن، رابطه/فضایی، و زمینه‌ای–رویدادی/چندمرحله‌ای)، نوع خروجی (ارزیابی تفکیکی در برابر تولیدی)، نوع تعامل (تک‌مرحله‌ای در برابر چندمرحله‌ای/دنباله‌ای)، ماهیتی (تصویر در برابر ویدئو)، و وجود سناریوهای ایمنی به‌همراه مدل تهدید صریح است. در کنار این‌ها، دو ستون نیز به‌صورت «محافظه‌کارانه» برای پوشش زبان/دامنه لحاظ شد (English-only و Domain-specific) که فقط در صورت تصریح مستقیم در معرفی داده/سنجه مقدار ۱ می‌گیرند؛ برای مثال، سنجه‌های مرتبط با متنِ صحنه یا پزشکی در دسته‌ی دامنه‌محور علامت می‌خورند.

همان‌گونه که در شکل ۶ مشاهده می‌شود، اکوسیستم ارزیابی فعلی معمولاً یا روی «توهم» متمرکز است یا روی «ایمنی/گریز امنیتی»، و پوشش همزمانِ توهم+ایمنی در یک پروتکل یکپارچه به‌ندرت دیده می‌شود. علاوه بر این، سنجه‌های چندمرحله‌ای (دیالوگ/ زنجیره تصویر/ ویدئو) هنوز نسبت به سنجه‌های تک‌تصویر کمتر و پراکنده‌تر هستند، در حالی که همین تنظیمات چندمرحله‌ای در عمل سطح خطا و سطح حمله‌ی بزرگ‌تری می‌سازند.

۵- استراتژی‌های کاهش خطر: زمان آموزش در مقابل زمان استنتاج

با توجه به اینکه توهم و شکست‌های ایمنی در مدل‌های بزرگ بینایی-زبان غالباً از «اتکای بیش از حد به پیش‌فرض‌های زبانی»، «ضعف گراندینگ بصری»، و «مکانیزم‌های تولید خودرگرسیو» نشأت می‌گیرند، ما راهبردهای کاهش را به دو طبقه‌ی عملیاتی تقسیم می‌کنیم: مداخلات زمان آموزش که توزیع رفتاری مدل را با داده/هدف بهینه‌سازی بازتعریف می‌کنند، و مداخلات زمان استنتاج که بدون بازآموزی سنگین، با کنترل مسیر تولید یا تصحیح/دفاع پسینی، ریسک پاسخ‌های نادرست یا ناایمن را کاهش می‌دهند [6], [26], [36].

از منظر استقرار، این تفکیک صرفاً طبقه‌بندی مفهومی نیست؛ بلکه تفاوت‌های معناداری در هزینه‌ی محاسباتی، نیازمندی داده، پیچیدگی پیاده‌سازی، و قابلیت بازگشت‌پذیری ایجاد می‌کند. به همین دلیل، در این بخش می‌کوشیم راهکارها را به‌صورت «قابل مقایسه و تصمیم‌ساز» ارائه کنیم، نه صرفاً فهرست مطالعات[6].



شکل ۶. نقشه حرارتی پوشش سنجه‌ها و پروتکل‌های ارزیابی توهم و ایمنی در مدل‌های بزرگ بینایی-زبان؛ مقدار ۱ نشان می‌دهد که بُعد مورد نظر به صراحت در تعریف یا پروتکل سنجه پوشش داده شده است.

۵-۱- مداخلات زمان آموزش

۵-۱-۱- داده محور: اصلاح سیگنال یادگیری از مسیر داده/ترجیحات

در مداخلات داده‌محور، فرض اصلی آن است که بخشی از توهم به‌علت «کیفیت ناکافی سیگنال آموزشی» (از جمله داده‌های دستورمحور، داده‌های ترجیحی، یا نویز کپشن) پایدار می‌ماند؛ بنابراین، اصلاح داده یا طراحی داده‌ی ترجیحی می‌تواند مستقیماً روی توزیع خروجی اثر بگذارد. در Factually Augmented RLHF، پاداش RLHF با اطلاعات «فکت‌محور» مانند کپشن تصویر و گزینه‌های چندگزینه‌ای تقویت می‌شود تا از پدیده‌ی reward hacking در RLHF کاسته شود و هم‌زمان هم‌ترازی چندوجهی بهبود یابد[36].

در خانواده‌ی DPO نیز، کیفیت و نحوه‌ی ساخت داده‌ی ترجیحی تعیین‌کننده معرفی می‌شود. OPA-DPO نشان می‌دهد اگر داده‌ی ترجیحی نسبت به policy مرجع «off-policy» باشد، شکاف توزیعی و نقش قید KL می‌تواند یادگیری را تضعیف کند؛ از این رو، چارچوبی پیشنهاد می‌شود که با استفاده از بازخورد خبره برای اصلاح پاسخ‌های توهمی، داده‌ی ترجیحی را به‌صورت on-policy هم‌راستا می‌سازد[38].

بازنویسی ترجیحات در سطح موضوع[1] نیز مسئله را از زاویه‌ی «پیکربندی شکاف پاداش» صورت‌بندی می‌کند و استدلال می‌کند موفقیت DPO به «reward gap واقعی» درون جفت‌های ترجیحی وابسته است؛ بنابراین بازنویسی ترجیحات در سطح موضوع برای کنترل نظام‌مند reward gap پیشنهاد می‌شود[33].

۵-۱-۲- هم ترازی: هم‌ترازی برای کاهش توهم و افزایش ایمنی

در محور کاهش توهم، RLHF و DPO هر دو تلاش می‌کنند «پاسخ وفادار به تصویر» را بر «پاسخ روان ولی غیرگراند» ترجیح دهند، اما تفاوت آنها عمدتاً در هزینه و کنترل‌پذیری است. Factually Augmented RLHF با تقویت مدل پاداش توسط سیگنال‌های فکت‌محور، ادعا می‌کند هم کیفیت هم‌ترازی و هم پایداری یادگیری را بهبود می‌دهد و اثر reward hacking را کاهش می‌دهد[36].

[1] TPR (Topic-level Preference Rewriting)

در سوی دیگر، V-DPO توهم را تا حدی پیامد «اتکای بیش از حد به LLM backbone و priorهای زبانی» می‌داند و با Vision-guided preference learning می‌کوشد توجه به متن را از priorهای زبانی جدا کرده و به سیگنال بصری بازگرداند؛ همچنین گزارش می‌کند که جفت‌های ترجیحی image-contrast می‌توانند برای استخراج ظرایف زمینه بصری مؤثرتر باشند. در محور ایمنی نیز، مطالعات نشان می‌دهند هم‌ترازی صرفاً در سطح LLM برای مدل‌های چندوجهی کافی نیست و لازم است ماهیت بصری به‌صورت صریح وارد سازوکار ایمنی شود[37]. Safety Alignment با افزودن ماژول‌های ایمنی (مانند safety projector/tokens/head) و آموزش دومرحله‌ای گزارش می‌کند که دفاع در برابر تصاویر پرریسک بهبود می‌یابد و حتی وجود محتوای بالقوه پرریسک در برخی دیتاست‌های چندوجهی رایج را مطرح می‌کند. PSA-VLM نیز با ایده‌ی concept bottleneck و هم‌ترازی تدریجی، هدف را افزایش کنترل‌پذیری و توضیح‌پذیری هم‌ترازی ایمنی قرار می‌دهد و ادعا می‌کند این کار می‌تواند دفاع در برابر تصاویر پرریسک را با اثر حداقلی بر عملکرد عمومی تقویت کند [4] ,[15]. مقایسه راهبردهای منتخب زمان آموزش از نظر هدف، سیگنال و داده موردنیاز، تغییرات مدل و هزینه آموزشی در جدول ۱۰ ارائه شده است.

جدول ۱۰. مقایسه روش‌های زمان آموزش

روش	هدف اصلی	سیگنال/داده‌ی موردنیاز	تغییر در مدل	هزینه‌ی آموزشی	نکته‌ی کلیدی)مطابق ادعای مقاله(
Factually Augmented RLHF [36]	کاهش توهم +بهبود هم‌ترازی چندوجهی	داده‌ی ترجیحی +پاداش تقویت‌شده با سیگنال‌های فکت‌محور (کپشن/گزینه‌ها)	نیازمند Reward Model و RLHF pipeline	بالا	کاهش reward hacking با افزودن اطلاعات فکت‌محور به پاداش.
V-DPO [37]	کاهش توهم با تقویت اتکا به تصویر	preference pairها (response-contrast و image-contrast)	fine-tuning با DPO	متوسط	کاهش over-reliance به priorهای زبانی و تقویت visual context learning
OPA-DPO [38]	کاهش توهم با داده‌ی on-policy	اصلاح پاسخ با بازخورد خبره + هم‌راستاسازی on-policy	fine-tuning با DPO	متوسط	نقش تعیین‌کننده‌ی on-policy بودن داده و تحلیل اثر KL نسبت به policy مرجع .
TPR [33]	بهینه‌سازی نظام‌مند preference	بازنویسی ترجیحات در سطح موضوع	سازگار با preference learning	متوسط	تأکید بر «reward gap configuration» به‌عنوان عامل کلیدی موفقیت DPO
Safety Alignment [4]	افزایش ایمنی ماهیت بصری	داده/برچسب‌های ایمنی +آموزش دومرحله‌ای	افزودن safety modules	متوسط	بهبود دفاع در برابر تصاویر پرریسک با ماژول‌های ایمنی اختصاصی.
PSA-VLM [15]	افزایش ایمنی با concept bottleneck	مفاهیم ایمنی +هم‌ترازی تدریجی	safety modules به‌صورت bottleneck	متوسط	افزایش کنترل‌پذیری/توضیح‌پذیری هم‌ترازی ایمنی

۵-۲- مداخلات زمان استنتاج

۵-۲-۱- رمزگشایی/تولید کنترل شده

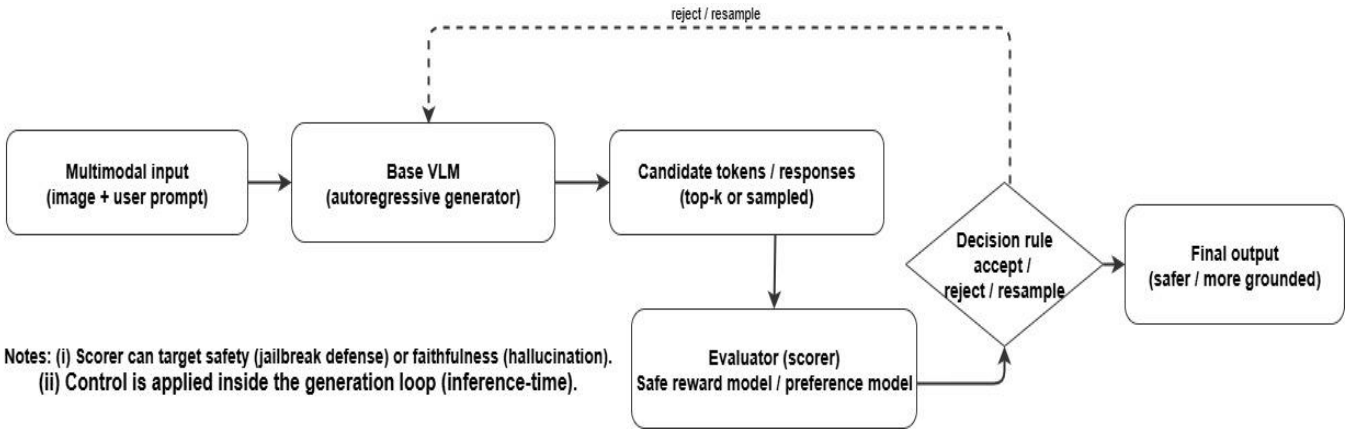
در مداخلات سطح رمزگشایی، ایده‌ی مشترک آن است که حتی بدون تغییر وزن‌ها، می‌توان مسیر تولید را به‌گونه‌ای هدایت کرد که احتمال تولید خروجی‌های توهمی یا ناایمن کاهش یابد. روش ایمیون نمونه‌ی شاخص این رویکرد در حوزه ایمنی است که از یک مدل پاداش ایمن در چارچوب رمزگشایی کنترل شده استفاده می‌کند و علاوه بر معرفی چارچوب، تحلیل نظری و نتایج تجربی برای کاهش نرخ موفقیت حمله ارائه می‌دهد [26].

در محور توهم، بهینه‌سازی ناشی از توهم[1] با تمرکز بر رمزگشایی کنتراست[2]، یک مدل ترجیح نظری را برای تشدید کنتراست میان توکن‌های هدف و توکن‌های توهمی وارد می‌کند و نشان می‌دهد که بهبود کیفیت «القای توکن‌های توهمی» می‌تواند کارایی رمزگشایی کنتراست را افزایش دهد [34].

DSCR نیز به‌عنوان یک روش training-free، اصلاحاتی را روی KV-cache اعمال می‌کند و با بهره‌گیری از سرنخ‌های عمق/فضایی، تلاش می‌کند توجه مدل را به نواحی مرتبط هدایت کند تا خروجی کمتر از تصویر جدا شود؛ این روش به‌طور صریح به‌عنوان مداخله‌ی زمان استنتاج و بدون نیاز به fine-tuning معرفی شده است.

[1] HIO (Hallucination-Induced Optimization)

[2] contrast decoding



شکل ۷. رمزگشایی کنترل شده/هم‌ترازی زمان استنتاج (جریان عمومی)

در شکل ۷، یک الگوی عمومی برای مداخلات زمان استنتاج ارائه شده است که در آن کنترل، مستقیماً داخل حلقه‌ی تولید اعمال می‌شود. در این الگو، تولید خودرگرسیو به‌جای آن‌که صرفاً بر اساس لگیت‌های مدل پایه پیش برود، با یک سیگنال ارزیابی ثانویه (ایمنی یا ترجیح/وفاداری) تنظیم می‌شود و تصمیم‌گیری در هر گام می‌تواند به پذیرش، رد یا بازنمونه‌گیری منجر شود. این سازوکار، به‌صورت طبیعی برای سناریوهای ایمنی مناسب است زیرا امکان «مهار مسیر تولید» را در مواجهه با ورودی‌های حمله فراهم می‌کند، و در سناریوهای توهم نیز می‌تواند با تشدید کنتراست میان توکن‌های هدف و توکن‌های توهمی، احتمال خروجی غیرگراند را کاهش دهد. از منظر مهندسی، این رویکرد یک trade-off روشن ایجاد می‌کند: افزایش کنترل‌پذیری در برابر افزایش سربار محاسباتی در زمان استنتاج، که باید متناسب با سطح ریسک کاربرد تنظیم شود.

۵-۲-۲- دفاع/تصحیح پس از تولید یا اصلاح درون‌مدلی

در دسته‌ی مداخلات پسینی، یک مسیر مهم «ویرایش بازنمایی‌های نهان» است. این روش با فرافکنی بازنمایی‌های تصویری به فضای واژگان و تحلیل میزان اطمینان روی اشیای واقعی در برابر اشیای توهمی، یک الگوریتم پاک‌سازی دانش پیشنهاد می‌کند که با متعامدسازی ویژگی‌های تصویر نسبت به ویژگی‌های توهمی، کاهش هدفمند توهم را بدون بازآموزی کامل دنبال می‌کند. [39].

در حوزه‌ی ایمنی، SafePTR یک دفاع token-level و training-free پیشنهاد می‌کند و گزارش می‌کند که سهم کوچکی از توکن‌ها در لایه‌های early-middle می‌تواند رفتار ناایمن را تحریک کند؛ از این رو، سازوکار prune-then-restore را برای حذف هدفمند توکن‌های مضر و بازیابی ویژگی‌های benign ارائه می‌دهد و ادعا می‌کند این کار بدون سربار محاسباتی اضافی انجام می‌شود[27].

۵-۳- تحلیل تطبیقی: هزینه، دقت، و قابلیت استقرار

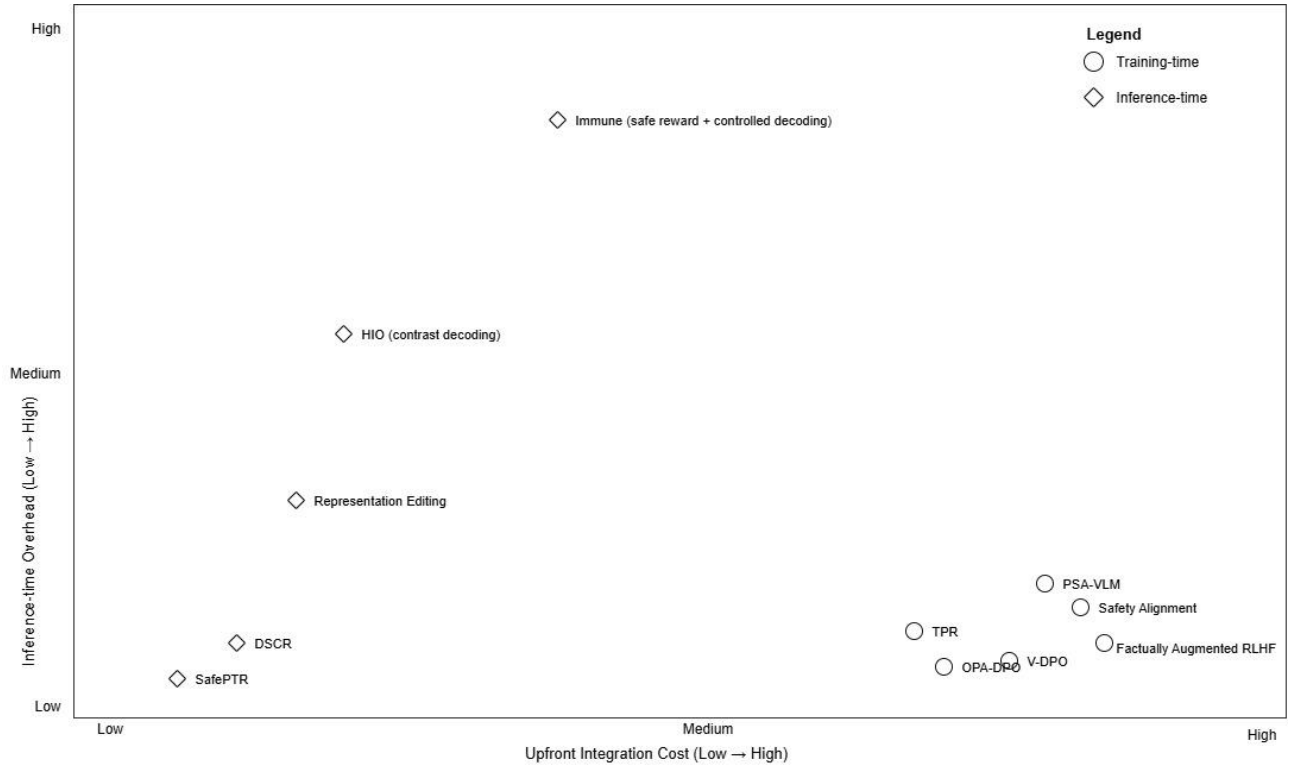
برای تصمیم‌گیری عملی، ما سه معیار را محور مقایسه قرار می‌دهیم: (۱) هزینه (داده/آموزش/محاسبه)، (۲) اثرگذاری (کاهش توهم یا کاهش موفقیت حمله طبق گزارش مقاله)، و (۳) قابلیت استقرار (نیاز به زیرساخت آموزش، تغییر معماری، یا سربار latency). بر این اساس، روش‌های training-time مانند RLHF/DPO معمولاً به داده‌ی ترجیحی و چرخه‌ی آموزشی نیاز دارند اما بهبود رفتاری پایدارتر ایجاد می‌کنند، در حالی که روش‌های inference-time یا training-free برای سخت‌سازی سریع محصول جذاب‌اند ولی ممکن است دامنه‌ی اثر یا پایداری آنها وابسته به معماری/تنظیمات تولید باشد[40] ,[36].

به‌طور مشخص، Immune با وارد کردن ارزیابی ایمنی در decoding، بهبود ایمنی را گزارش می‌کند اما ذاتاً می‌تواند سربار محاسباتی ایجاد کند، زیرا امتیازدهی/کنترل در حلقه‌ی تولید انجام می‌شود[26].

در مقابل، SafePTR صراحتاً training-free بودن و عدم سربار اضافی را مطرح می‌کند و DSCR نیز بدون fine-tuning معرفی شده است؛ از این رو، این دو روش از منظر استقرار سریع و محدودیت latency معمولاً مناسب‌ترند، هرچند میزان و شرایط اثرگذاری باید مطابق پروتکل‌های ارزیابی گزارش‌شده در هر مقاله سنجیده شود[40]. مقایسه روش‌های زمان استنتاج از نظر هدف، مؤلفه‌های اضافی، سربار محاسباتی و ملاحظات استقرار در جدول ۱۱ ارائه شده است.

جدول ۱۱. تریدآف‌های زمان استنتاج (تأخیر در مقابل دقت)

روش	هدف	نیاز به مؤلفه‌ی اضافی	سربار زمان استنتاج (کیفی)	نکته‌ی استقراری (طبق مقاله)
Immune	کاهش jailbreak/ناایمنی	safe reward model + controlled decoding	متوسط تا زیاد	دفاع زمان استنتاج با تضمین/تحلیل نظری و ارزیابی روی سنجه‌ها.
SafePTR	دفاع token-levelدر برابر jailbreak	خیر(training-free)	بسیار کم	prune-then-restore برای حذف توکن‌های مضر، بدون سربار اضافی گزارش‌شده.
DSCR	کاهش توهم با گراندینگ فضایی/عمقی	خیر (training-free)	کم	اصلاح KV-cache با سرنخ‌های فضایی/عمق و بدون fine-tuning
HIO	کاهش توهم در contrast decoding	نظری preference model	متوسط	تقویت contrast decoding با تشدید کنتراست میان توکن‌های هدف و توهمی.
Representation Editing	کاهش توهم با ویرایش بازنمایی	رویه‌ی ویرایش نهان	کم تا متوسط	ویرایش هدفمند ویژگی‌های تصویر برای حذف توهم با حفظ عملکرد گزارش‌شده.



شکل ۱. کاهش زمان آموزش در مقابل زمان استنتاج (پراکندگی کیفی)

در شکل ۸، ما این نگاشت را به‌عنوان یک چارچوب کیفی و تصمیم‌محور تفسیر می‌کنیم که هدف آن کمی‌سازی نیست، بلکه مقایسه‌ی عملیاتی راهبردهای کاهش در دو محور «هزینه/پیچیدگی بالادستی» و «سربار زمان استنتاج» است. هر نقطه صرفاً بر اساس تعریف روش در خود مقاله (نیاز به داده‌ی ترجیحی و چرخه‌ی آموزش، تغییرات معماری، یا مداخله در حلقه‌ی decoding) در یکی از نواحی Low/Medium/High قرار می‌گیرد؛ بنابراین، این شکل یک نقشه‌ی قابل تکرار برای طبقه‌بندی مداخلات است، نه یک ادعای عددی درباره‌ی برتری مطلق روش‌ها[6], [26], [36].

بر این اساس، خوشه‌ی مداخلات زمان آموزش (نظیر RLHF/DPO و هم‌خانواده‌ها) عمدتاً در ناحیه‌ی «هزینه‌ی بالادستی بالاتر و سربار استنتاجی پایین‌تر» قرار می‌گیرد، زیرا بار اصلی آن‌ها در مرحله‌ی گردآوری داده‌ی ترجیحی/هم‌ترازی و آموزش اعمال می‌شود. در مقابل، بخشی از مداخلات زمان استنتاج که ماهیت training-free یا سبک دارند (مانند دفاع‌های token-level یا اصلاح KV-cache) در ناحیه‌ی «هزینه‌ی بالادستی پایین و سربار استنتاجی پایین» قرار

می‌گیرند. همچنین، روش‌های مبتنی بر controlled decoding با مدل پاداش/ترجیح به‌طور طبیعی به سمت «سربار استنتاجی بالاتر» متمایل می‌شوند، زیرا کنترل ایمنی/وفاداری در حلقه‌ی تولید اعمال می‌گردد. در نهایت، رویکردهایی مانند ویرایش بازنمایی به‌عنوان مداخله‌ی میانی نمایش داده می‌شوند که با حداقل‌سازی بازآموزی، تغییرات هدفمند را از مسیر تحلیل/دستکاری نهان دنبال می‌کنند[34] ,[26] ,[26] ,[7].

۶- بحث

با وجود رشد سریع توانمندی‌های مدل‌های بزرگ بینایی-زبان، شکاف اعتماد همچنان پابرجاست زیرا «کیفیت تولید» لزوماً به «پایبندی به شواهد بصری» منجر نمی‌شود و در بسیاری از سناریوها، مدل می‌تواند پاسخ‌های روان اما غیرگراند تولید کند. شواهد سنجه‌ی نشان می‌دهند که حتی در ارزیابی‌های هدفمندِ توهم شیء، سوگیری‌های پاسخ (از جمله تمایل به پاسخ مثبت) و حساسیت به قالب پرسش/دستور می‌تواند نرخ توهم مشاهده‌شده را تحت تأثیر قرار دهد و تفسیر نتایج را دشوار کند. عامل دوم، «بحران اکوسیستم ارزیابی» است: بخشی از سنجه‌ها ماهیت تفکیکی دارند، بخشی تولیدی‌اند، و بخشی نیز چندمرحله‌ای یا مبتنی بر ورودی‌های پیچیده‌تر؛ بنابراین مدل ممکن است در یک پروتکل خوب و در دیگری شکننده باشد. برای نمونه، HallusionBench به‌طور صریح به پیچیدگی توهم‌های درهم‌تنیده (زبان/بینایی) و خطاهای ناشی از استدلال یا دانش پیشین اشاره می‌کند و نشان می‌دهد ارزیابی‌های ساده نمی‌توانند تمام طیف شکست را پوشش دهند. عامل سوم، شکاف میان «هم‌ترازی ایمنی» و «پایداری در برابر ورودی‌های چندوجهی» است. نتایج سنجه‌های ایمنی نشان می‌دهد تصاویر «مرتبط با پرسش» می‌توانند همانند ورودی‌های متنی مخرب عمل کنند و مکانیزم‌های هم‌ترازیِ صرفاً زبانی را دور بزنند. در ادامه، مطالعات دفاعی نیز تصریح می‌کنند که هم‌ترازی زمان آموزش به تنهایی کافی نیست و برای بستن شکاف ایمنی باید به مداخلات زمان استنتاج یا کنترل‌های اختصاصی چندوجهی نیز توجه کرد.

در نهایت، یک توضیح ریشه‌ای که به «ماندگاری» شکاف اعتماد کمک می‌کند، پدیده‌هایی نظیر شکاف ماهیتی و عدم‌تعادل اطلاعاتی در یادگیری بازنمایی‌های مشترک است؛ یعنی حتی پیش از مرحله‌ی تولید، هم‌ترازی هندسی و محتوایی میان سیگنال‌های متن و تصویر می‌تواند ناقص باشد و زمینه را برای اتکای بیش از حد به priorهای زبانی فراهم کند. مسیر طبیعیِ توسعه‌ی VLMها از «تصویر منفرد» به «ویدئو، توالی فریم‌ها و تعاملات چندمرحله‌ای» حرکت می‌کند؛ در این فضا، مسئله‌ی توهم فقط به "وجود/عدم وجود یک شیء" محدود نیست، بلکه به پایداری بین فریم‌ها، سازگاری زمانی، و یکپارچگی روایت گره می‌خورد. در چارچوب‌های اخیرِ حمله، ویدئو نه‌تنها سطح حمله را بزرگ‌تر می‌کند بلکه امکان تزریق تدریجی نشانه‌ها و ایجاد تعارضات بین‌فریمی را فراهم می‌سازد؛ به‌طور مشخص، VideoJail نشان می‌دهد ماهیت ویدئو می‌تواند آسیب‌پذیری‌های هم‌ترازی را تشدید کند و دفاع‌های موجود را به چالش بکشد.

از منظر پژوهشی، ما پیشنهاد می‌کنیم «توهم زمانی/زمینه‌ای» به‌عنوان یک مسئله‌ی مستقل مدل‌سازی شود: (۱) توهم گذرا (ظاهر/ناپدید شدن غیرواقعی اشیاء)، (۲) drift روایی (انحراف تدریجی پاسخ در تعامل چندمرحله‌ای)، و (۳) ناسازگاری سببی (ادعاهای علت-معلولی که با توالی فریم‌ها سازگار نیست). چنین صورت‌بندی‌ای کمک می‌کند تا ارزیابی از سطح تک‌تصویر به سمت معیارهای پایداری و سازگاری توالی حرکت کند و امکان طراحی پروتکل‌های استانداردتر فراهم شود.

در سطح روش، جهت‌های آینده برای کاهش توهم زمانی به‌طور طبیعی شامل قیدهای سازگاری بین‌فریمی (Consistency Constraints)، حافظه‌ی ساختاریافته برای اشیاء و روابط در زمان، و اعتبارسنجی چندمرحله‌ای است؛ همچنین، با توجه به اینکه ویدئو در کنار فرصت، تهدید نیز ایجاد می‌کند، لازم است ارزیابی ایمنی به سمت «مدل تهدید توالی‌محور» و سنجش انتقال‌پذیری حملات از تصویر به ویدئو حرکت کند.

توضیح‌پذیری در VLMها صرفاً یک ویژگی تزئینی نیست، بلکه یک ابزار برای کاهش ریسک عملیاتی است: اگر نتوانیم تشخیص دهیم مدل چرا به یک خروجی رسیده، نمی‌توانیم میان «خطای ادراکی»، «اتکای بیش از حد به prior زبانی»، یا «انحراف ناشی از حمله» تمایز بگذاریم. یکی از مسیرهای امیدوارکننده، تحلیل و ویرایش بازنمایی‌های نهان است؛ برای نمونه، Representation Editing نشان می‌دهد می‌توان بازنمایی‌های داخلی را به فضای واژگان فرافکنی کرد، تفاوت اعتماد بین اشیای واقعی و توهمی را مشاهده کرد، و با ویرایش هدفمند ویژگی‌ها، نرخ توهم را کاهش داد.

از سوی دیگر، تحلیل‌های ایمنیِ مبتنی بر سازوکارهای درونی نشان می‌دهند که «سطح حمله» در MLLMها می‌تواند به زیرساخت‌های توکنی/لایه‌ای گره بخورد؛ برای مثال، SafePTR با تحلیل توکن‌محور گزارش می‌کند بخش بسیار کوچکی از توکن‌ها در برخی لایه‌ها می‌تواند رفتار ناامن را تحریک کند و این مشاهده، مسیر دفاع‌های دقیق‌تر و کم‌هزینه‌تر را باز می‌کند.

در کنار این‌ها، توضیح‌پذیری باید با «گراندینگ قابل آزمون» همراه شود؛ یعنی به‌جای اتکا به توضیحات زبانی، شواهدی مانند نواحی توجه/پشتیبان، سازگاری با اشیاء موجود، یا قواعد سازگاری زمانی ارائه گردد. تحلیل‌های ریشه‌ای درباره شکاف ماهیتی نیز به‌عنوان ابزار تشخیصی مهم‌اند، زیرا نشان می‌دهند مشکل می‌تواند قبل از مرحله‌ی استدلال، در خود فضای بازنمایی مشترک شکل بگیرد.

در سطح پژوهشی، ما توصیه می‌کنیم حرکت از «بهینه‌سازی یک معیار واحد» به سمت طراحی پروتکل‌های ارزیابی چندبعدی تسریع شود؛ به‌ویژه، سنجه‌هایی که هم‌زمان توهم (در سطوح شیء/ویژگی/رابطه/زمینه) و ایمنی (مدل تهدید و سناریوهای آسیب) را پوشش می‌دهند، می‌توانند از پراکندگی اکوسیستم ارزیابی بکاهند و مانع بهینه‌سازی‌های تک‌بعدی شوند. در سطح صنعت، رویکردهای دفاعی باید به‌صورت «چندلایه» طراحی شوند: (۱) هم‌ترازی/بهبود رفتار در زمان آموزش، (۲) کنترل و اعتبارسنجی زمان استنتاج برای سناریوهای پرریسک، و (۳) پایش پس از استقرار با سنجه‌های مرتبط با دامنه‌ی کاربرد. شواهد موجود نشان می‌دهد دفاع‌های زمان استنتاج مبتنی بر کنترل decoding می‌توانند شکاف ایمنی را در برابر گریز امنیتی‌ها کاهش دهند، اما باید هم‌زمان هزینه‌ی latency و ریسک افت کارایی نیز در نظر گرفته شود. در سطح سیاست‌گذاری و حکمرانی، تمرکز صرف بر «خروجی‌های نامطلوب» کافی نیست و لازم است الزامات حداقلی برای شفافیت داده/پروتکل ارزیابی، گزارش‌پذیری رخدادهای ایمنی، و ممیزی مبتنی بر سناریو تعریف شود؛ به‌ویژه با ورود ویدئو، سطح حمله و دامنه‌ی پیامدها افزایش می‌یابد و نیاز به استانداردسازی مدل تهدید و گزارش‌دهی حملات پررنگ‌تر می‌شود.

جدول ۱۲. توصیه‌های عملی

مخاطب	توصیه عملیاتی	دلیل/شاهد در ادبیات
پژوهشگر	طراحی ارزیابی چندبعدی)توهم +ایمنی (به‌جای معیار تک‌بعدی	پراکندگی و ناهمگنی پروتکل‌ها می‌تواند نتایج را غیرقابل مقایسه کند. [15] ,[1]
پژوهشگر	افزودن سنجه‌های پایداری زمانی و سازگاری بین‌فریمی برای ویدئو	ویدئو سطح حمله/شکست جدید ایجاد می‌کند و توالی‌محور است. [24]
پژوهشگر	تمرکز بر تحلیل بازنمایی و ابزارهای تشخیصی قابل آزمون	ویرایش/تحلیل بازنمایی می‌تواند مسیر کاهش هدفمند توهم فرا هم کند [39].
صنعت	دفاع چندلایه :آموزش‌محور +کنترل استنتاج + پایش پس از استقرار	هم‌ترازی زمان آموزش به‌تنهایی کافی نیست و کنترل decoding می‌تواند ریسک را کاهش دهد.[26]
صنعت	تفکیک سناریوهای پرریسک و فعال‌سازی گیت ایمنی/اعتبارسنجی در آن‌ها	تصاویر مرتبط می‌توانند هم‌ترازی را بشکنند و نیاز به گیت پررنگ است [14].
صنعت	مستندسازی مدل تهدید و سنجه‌های عملیاتی (latency/utility) هنگام دفاع	کنترل decodingمی‌تواند سربار ایجاد کند و باید در تصمیم استقرار لحاظ شود .[26].
سیاست‌گذار	الزام حداقلی برای گزارش‌پذیری رخدادهای ایمنی و ممیزی سناریومحور	با گسترش ویدئو، سطح حمله و پیامدها افزایش می‌یابد .[24].
سیاست‌گذار	تشویق به شفافیت داده/پروتکل ارزیابی و انتشار نتایج قابل بازتولید	ناهمگنی ارزیابی‌ها و شکاف اعتماد با شفافیت کاهش می‌یابد .[1].

جدول ۱۲ را به‌عنوان یک «خروجی تصمیم‌محور» تفسیر می‌کنیم که هدف آن فشرده‌سازی پیامدهای عملی مرور، و تبدیل یافته‌های پراکنده به مجموعه‌ای از اقدامات قابل اجرا برای سه گروه ذی‌نفع (پژوهشگران، صنعت، و سیاست‌گذاران) است. این نوع جمع‌بندی در بخش بحث، به‌جای تکرار نتایج، بر دلالت‌ها و توصیه‌های صریح تمرکز می‌کند و به خواننده اجازه می‌دهد مسیرهای اقدام را مستقیماً از ادبیات استخراج‌شده دنبال کند. همچنین، توصیه‌های جدول را «نسخه نهایی و جهان‌شمول» تلقی نمی‌کنیم؛ بلکه آن‌ها را به‌صورت مجموعه‌ای از گزاره‌های قابل انطباق ارائه می‌کنیم که باید با زمینه‌ی استقرار (دامنه کاربرد، حساسیت ریسک، قیود تأخیر/هزینه، و مدل تهدید) کالیبره شوند. از این منظر، جدول ۱۲ هم‌زمان دو کارکرد دارد: (۱) برجسته‌سازی خلأهای پژوهشی که مانع بسته‌شدن شکاف اعتماد هستند، و (۲) ارائه‌ی حداقل اقدامات مهندسی/حاکمیتی که می‌توانند بدون وابستگی به یک مدل خاص، ریسک توهم و شکست ایمنی را کاهش دهند.

۷- نتیجه‌گیری

در این مرور نظام‌مند، ادبیات مرتبط با قابلیت اعتماد در مدل‌های بزرگ بینایی-زبان را با تمرکز بر دو محور کلیدی «توهم» و «ایمنی» یکپارچه‌سازی کردیم و نشان دادیم که خطاهای این مدل‌ها صرفاً به یک نوع وظیفه یا یک پروتکل ارزیابی محدود نمی‌شوند، بلکه به‌صورت الگوهای تکرارشونده در طیفی از سناریوها ظاهر می‌گردند. جمع‌بندی شواهد نشان می‌دهد توهم می‌تواند در سطوح مختلفی از توصیف شیء تا ویژگی‌ها، روابط و زمینه رخ دهد و در تنظیمات چندمرحله‌ای یا زمینه‌محور تشدید شود؛ هم‌زمان، ایمنی نیز به‌ویژه در مواجهه با ورودی‌های چندوجهی، شکننده است و می‌تواند تحت شرایطی که مدل از منظر زبانی هم‌تراز به‌نظر می‌رسد، دچار شکست عملیاتی شود.

یافته‌ی محوری این مطالعه آن است که « قابلیت اعتماد گلوگاه » در VLMها از شکاف میان توانایی تولید زبانیِ روان و الزام به تولید مبتنی بر شواهد و قابل راستی‌آزمایی ناشی می‌شود. به بیان دیگر، پیشرفت در قابلیت‌های ادراکی یا افزایش اندازه و مقیاس مدل، لزوماً به افزایش قابلیت اعتماد منجر نمی‌شود، مگر آن‌که سازوکارهای یادگیری و ارزیابی به‌صورت صریح، مدل را به وفاداری به شواهد بصری و پاسخ‌گویی ایمن مقید کنند. از این منظر، مسئله‌ی قابلیت اعتماد[1] صرفاً «بهبود دقت» نیست، بلکه «مهار رفتار تولیدی» تحت قیود سازگاری درونی، سازگاری بیرونی و الزامات ایمنی در شرایط واقعی است.

از منظر روش‌شناسی، ما نشان دادیم اکوسیستم ارزیابی موجود به‌دلیل ناهمگنی در قالب خروجی (تفکیکی در برابر تولیدی)، نوع تعامل (تک‌مرحله‌ای در برابر چندمرحله‌ای) و تفاوت در تعریف عملیاتی توهم و ایمنی، ریسک نتیجه‌گیری‌های غیرقابل تعمیم را افزایش می‌دهد. همین پراکندگی باعث می‌شود یک مدل در یک سنجه «قابل اتکا» و در تنظیمات نزدیک به کاربرد واقعی «شکننده» باشد. بنابراین، بحران اندازه‌گیری نه یک مسئله‌ی جانبی، بلکه بخشی از منشأ شکاف اعتماد است.

بر مبنای این نتایج، چند جهت‌گیری مشخص برای افزایش قابلیت اعتماد پیشنهاد می‌شود. نخست، توسعه‌ی ارزیابی یکپارچه که توهم (در سطوح شیء/ویژگی/رابطه/زمینه-زمان) و ایمنی (با مدل تهدید روشن و سناریوهای صریح) را در یک پروتکل بازتولیدپذیر گرد هم آورد و امکان مقایسه‌ی منصفانه‌ی روش‌ها را فراهم کند. دوم، حرکت از ارزیابی تک‌تصویر به سمت ارزیابی‌های چندمرحله‌ای و زمانی ضروری است، زیرا بسیاری از شکست‌های مهم در تعامل، حافظه‌ی گفتگو و پایداری زمانی بروز می‌کنند. سوم، در سطح طراحی سیستم، رویکرد safety-first باید به‌عنوان یک اصل معماری اتخاذ شود؛ یعنی دفاع چندلایه شامل بهبودهای زمان آموزش، کنترل‌ها و اعتبارسنجی‌های زمان استنتاج، و پایش پس از استقرار در کنار هم قرار گیرند، نه به‌صورت وصله‌های دیرهنگام.

در نهایت، پیام این مقاله آن است که ارزش عملی VLMها در کاربردهای حساس تنها زمانی محقق می‌شود که «توانمندی» به «قابلیت اعتماد قابل سنجش» و سپس به «ایمنی قابل اتکا در استقرار» تبدیل شود. مسیر آینده، نه صرفاً در بزرگ‌تر کردن مدل‌ها، بلکه در استانداردسازی ارزیابی، غنی‌سازی سناریوهای واقع‌گرایانه، و طراحی سیستم‌های چندوجهی با اولویت ایمنی و راستی‌آزمایی نهفته است.

منابع

[1] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, D. Manocha, and T. Zhou, "HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models," Mar. 25, 2024, arXiv:2310.14566. doi: 10.48550/arXiv.2310.14566.

[2] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating Object Hallucination in Large Vision-Language Models," Oct. 26, 2023, arXiv: arXiv:2305.10355. doi: 10.48550/arXiv.2305.10355.

[3] A. Seth, D. Manocha, and C. Agarwal, "Towards a Systematic Evaluation of Hallucinations in Large-Vision Language Models," Mar. 13, 2025, arXiv: arXiv:2412.20622. doi: 10.48550/arXiv.2412.20622.

[4] Z. Liu, Y. Nie, Y. Tan, X. Yue, Q. Cui, C. Wang, X. Zhu, and B. Zheng, "Safety Alignment for Vision Language Models," May 22, 2024, arXiv:2405.13581. doi: 10.48550/arXiv.2405.13581.

[5] W. Luo, S. Ma, X. Liu, X. Guo, and C. Xiao, "JailBreakV: A Benchmark for Assessing the Robustness of MultiModal Large Language Models against Jailbreak Attacks," Nov. 24, 2024, arXiv: arXiv:2404.03027. doi: 10.48550/arXiv.2404.03027.

[6] Z. Yang, J. Fan, A. Yan, E. Gao, X. Lin, T. Li, K. Mo, and C. Dong, "Distraction is All You Need for Multimodal Large Language Model Jailbreaking," Jun. 17, 2025, arXiv:2502.10794. doi: 10.48550/arXiv.2502.10794.

[7] C. Jiang, H. Jia, W. Ye, M. Dong, H. Xu, M. Yan, J. Zhang, and S. Zhang, "Hal-Eval: A Universal and Fine-grained Hallucination Evaluation Framework for Large Vision Language Models," in Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia: ACM, Oct. 2024, pp. 525–534. doi: 10.1145/3664647.3680576.

[8] Q. Cao, J. Cheng, X. Liang, and L. Lin, "VisDiaHalBench: A Visual Dialogue Benchmark For Diagnosing Hallucination in Large Vision-Language Models," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 12161–12176. doi: 10.18653/v1/2024.acl-long.658.

[9] S. Schrodi, D. T. Hoffmann, M. Argus, V. Fischer, and T. Brox, "Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models," Apr. 16, 2025, arXiv: arXiv:2404.07983. doi: 10.48550/arXiv.2404.07983.

[10] S. Yin, C. Fu, S. Zhao, T. Xu, H. Wang, D. Sui, Y. Shen, K. Li, X. Sun, and E. Chen, "Woodpecker: Hallucination Correction for Multimodal Large Language Models," Science China Information Sciences, vol. 67, no. 12, p. 220105, Dec. 2024. doi: 10.1007/s11432-024-4251-x.

[11] V. Rawte, A. Mishra, A. Sheth, and A. Das, "Defining and quantifying visual hallucinations in vision-language models," in Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025), 2025, pp. 501–510.

[12] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, "Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning," Oct. 19, 2022, arXiv: arXiv:2203.02053. doi: 10.48550/arXiv.2203.02053.

[13] C. Yi, Y.-H. He, D.-C. Zhan, and H.-J. Ye, "Bridge the Modality and Capability Gaps in Vision-Language Model Selection," May

[1] trustworthiness

LLMs via Inference-Time Alignment," Jun. 14, 2025, arXiv:2411.18688. doi: 10.48550/arXiv.2411.18688.

[27] B. Chen, X. Lyu, L. Gao, J. Song, and H. T. Shen, "SafePTR: Token-Level Jailbreak Defense in Multimodal LLMs via Prune-then-Restore Mechanism," Dec. 03, 2025, arXiv: arXiv:2507.01513. doi: 10.48550/arXiv.2507.01513.

[28] W. You, Z. Li, H. Guo, X. Liu, W. Wang, X. Dong, and G. Wang, "MIRAGE: Multimodal Immersive Reasoning and Guided Exploration for Red-Team Jailbreak Attacks," Mar. 24, 2025, arXiv:2503.19134. doi: 10.48550/arXiv.2503.19134.

[29] L. Zhu, D. Ji, T. Chen, P. Xu, J. Ye, and J. Liu, "IBD: Alleviating Hallucinations in Large Vision-Language Models via Image-Biased Decoding," 2024, arXiv:2402.18476. doi: 10.48550/arXiv.2402.18476.

[30] Y. Shu, H. Lin, Y. Liu, Y. Zhang, G. Zeng, Y. Li, Y. Zhou, S.-N. Lim, H. Yang, and N. Sebe, "When Semantics Mislead Vision: Mitigating Large Multimodal Models Hallucinations in Scene Text Spotting and Understanding," Oct. 07, 2025, arXiv:2506.05551. doi: 10.48550/arXiv.2506.05551.

[31] X. Liu, M. Luo, A. Chatterjee, H. Wei, C. Baral, and Y. Yang, "Investigating VLM Hallucination from a Cognitive Psychology Perspective: A First Step Toward Interpretation with Intriguing Observations," arXiv preprint arXiv:2507.03123, 2025.

[32] E. Shayegani, Y. Dong, and N. Abu-Ghazaleh, "Jailbreak in Pieces: Compositional Adversarial Attacks on Multi-Modal Language Models," Oct. 10, 2023, arXiv:2307.14539. doi: 10.48550/arXiv.2307.14539.

[33] L. He, Z. Chen, Z. Shi, T. Yu, J. Shao, and L. Sheng, "Systematic Reward Gap Optimization for Mitigating VLM Hallucinations," Nov. 24, 2025, arXiv:2411.17265. doi: 10.48550/arXiv.2411.17265.

[34] X. Lyu, B. Chen, L. Gao, J. Song, and H. T. Shen, "Alleviating Hallucinations in Large Vision-Language Models through Hallucination-Induced Optimization," Apr. 01, 2025, arXiv: arXiv:2405.15356. doi: 10.48550/arXiv.2405.15356.

[35] Y. Shu, H. Lin, Y. Liu, Y. Zhang, G. Zeng, Y. Li, Y. Zhou, S.-N. Lim, H. Yang, and N. Sebe, "TextHalu-Bench and Semantic Hallucination Mitigation for Scene Text Understanding in Large Multimodal Models," 2025, arXiv:2506.05551. doi: 10.48550/arXiv.2506.05551.

[36] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, K. Keutzer, and T. Darrell, "Aligning Large Multimodal Models with Factually Augmented RLHF," Sep. 25, 2023, arXiv:2309.14525. doi: 10.48550/arXiv.2309.14525.

[37] Y. Xie, G. Li, X. Xu, and M.-Y. Kan, "V-DPO: Mitigating Hallucination in Large Vision Language Models via Vision-Guided Direct Preference Optimization," Nov. 05, 2024, arXiv: arXiv:2411.02712. doi: 10.48550/arXiv.2411.02712.

[38] Z. Yang, X. Luo, D. Han, Y. Xu, and D. Li, "Mitigating Hallucinations in Large Vision-Language Models via DPO: On-Policy Data Hold the Key," Mar. 03, 2025, arXiv: arXiv:2501.09695. doi: 10.48550/arXiv.2501.09695.

[39] N. Jiang, A. Kachinthaya, S. Petryk, and Y. Gandelsman, "Interpreting and Editing Vision-Language Representations to Mitigate Hallucinations," Feb. 10, 2025, arXiv: arXiv:2410.02762. doi: 10.48550/arXiv.2410.02762.

[40] X. Chen, Z. Ma, X. Zhang, S. Xu, S. Qian, J. Yang, D. F. Fouhey, and J. Chai, "Multi-Object Hallucination in Vision-Language Models," Oct. 31, 2024, arXiv:2407.06192. doi: 10.48550/arXiv.2407.06192.

18, 2025, arXiv: arXiv:2403.13797. doi: 10.48550/arXiv.2403.13797.

[14] X. Liu, Y. Zhu, J. Gu, Y. Lan, C. Yang, and Y. Qiao, "MM-SafetyBench: A Benchmark for Safety Evaluation of Multimodal Large Language Models," Jun. 19, 2024, arXiv: arXiv:2311.17600. doi: 10.48550/arXiv.2311.17600.

[15] Z. Liu, Y. Nie, Y. Tan, J. Liu, X. Yue, Q. Cui, C. Wang, X. Zhu, and B. Zheng, "PSA-VLM: Enhancing Vision-Language Model Safety through Progressive Concept-Bottleneck-Driven Alignment," 2024, arXiv:2411.11543. doi: 10.48550/arXiv.2411.11543.

[16] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," 2023, arXiv:2311.05232. doi: 10.48550/arXiv.2311.05232.

[17] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou, "Hallucination of Multimodal Large Language Models: A Survey," Apr. 01, 2025, arXiv:2404.18930. doi: 10.48550/arXiv.2404.18930.

[18] Z. Yang, X. Luo, D. Han, Y. Xu, and D. Li, "Mitigating Hallucinations in Large Vision-Language Models via DPO: On-Policy Data Hold the Key," 2025, arXiv. doi: 10.48550/ARXIV.2501.09695.

[19] P. Kaul, Z. Li, H. Yang, Y. Dukler, A. Swaminathan, C. J. Taylor, and S. Soatto, "THRONE: An Object-based Hallucination Benchmark for the Free-form Generations of Large Vision-Language Models," Apr. 03, 2025, arXiv:2405.05256. doi: 10.48550/arXiv.2405.05256.

[20] X. Fan, X. Wang, H. Wei, X. Zhang, and D. Zhao, "Toward a Stable, Fair, and Comprehensive Evaluation of Object Hallucination in Large Vision-Language Models," in Advances in Neural Information Processing Systems 37, Vancouver, BC, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024, pp. 111406–111431. doi: 10.52202/079017-3538.

[21] M. Ye-Bin, N. Hyeon-Woo, W. Choi, and T.-H. Oh, "BEAF: Observing BEfore-AFter Changes to Evaluate Hallucination in Vision-language Models," Jul. 18, 2024, arXiv: arXiv:2407.13442. doi: 10.48550/arXiv.2407.13442.

[22] M. Ye, X. Rong, W. Huang, B. Du, N. Yu, and D. Tao, "A Survey of Safety on Large Vision-Language Models: Attacks, Defenses and Evaluations," Feb. 14, 2025, arXiv: arXiv:2502.14881. doi: 10.48550/arXiv.2502.14881.

[23] E. Shayegani, Y. Dong, and N. Abu-Ghazaleh, "Jailbreak in pieces: Compositional Adversarial Attacks on Multi-Modal Language Models," Oct. 10, 2023, arXiv: arXiv:2307.14539. doi: 10.48550/arXiv.2307.14539.

[24] W. Hu, S. Gu, Y. Wang, and R. Hong, "VideoJail: Exploiting Video-Modality Vulnerabilities for Jailbreak Attacks on Multimodal Large Language Models," presented at the ICLR 2025 Workshop on Building Trust in Language Models and Applications, Mar. 2025. Accessed: Feb. 03, 2026. [Online]. Available: https://openreview.net/forum?id=fSAIDcPduZ

[25] H. Zhong, Q. Teng, B. Zheng, G. Chen, Y. Tan, Z. Liu, J. Liu, W. Su, X. Zhu, B. Zheng, and K. Zhang, "Towards Visualization-of-Thought Jailbreak Attack against Large Visual Language Models," presented at the Thirty-ninth Annual Conference on Neural Information Processing Systems, Oct. 2025. Accessed: Feb. 03, 2026. [Online]. Available: https://openreview.net/forum?id=5P5YgohyBZ.

[26] S. S. Ghosal, S. Chakraborty, V. Singh, T. Guan, M. Wang, A. Velasquez, A. Beirami, F. Huang, D. Manocha, and A. S. Bedi, "Immune: Improving Safety Against Jailbreaks in Multi-modal