ARE IMAGE-TO-VIDEO GENERATION MODELS ROBUST TO 3D VISUAL ILLUSIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-to-video generation models have shown strong ability to produce visually realistic videos with coherent motion and dynamic scene evolution from a single input image. However, their robustness under real-world 3D visual illusions remains unclear: in such cases, the first-frame appearance may conflict with the observable spatial, optical, projection, surface, object-identity, or lighting relations. Camera or object motion should remain compatible with the observable geometric, optical, and structural constraints, rather than preserving the illusion. Existing video generation benchmarks do not systematically evaluate failures in this illusion setting. To systematically evaluate this underexplored setting, we introduce VIVG-Bench (Visual Illusion Video Generation Benchmark), a benchmark of 680 carefully curated image-to-video prompts spanning diverse real-world 3D visual illusions. Each prompt is written without illusion descriptions: it removes explicit descriptions of the illusion mechanism and specifies only the intended motion or interaction, avoiding information leakage and requiring models to maintain physical consistency from the observable evidence in the input scene. We evaluate state-of-the-art video generation models, including both closed-source and open-source ones, using a human-centered protocol based on semantic/action adherence, hidden-relation consistency, and video stability. We define full success as following the requested motion, remaining consistent with observable geometric, optical, and structural constraints, and remaining visually stable. Under this criterion, even the best-performing model achieves only a 65.01% full-success rate. These results show that visually stable videos in this setting often remain incompatible with observable constraints, and VIVG-Bench provides a diagnostic testbed for evaluating and improving reliable video generation.

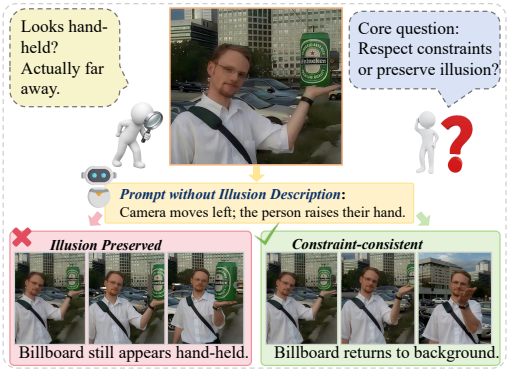


Figure 1: **Motivation of VIVG-Bench.** Given an illusion image, an image-to-video model should generate a video compatible with the observable geometric, optical, and structural constraints when the viewpoint changes or an object interaction occurs. This tests whether, during video generation, the model remains consistent with the observable spatial, optical, projection, surface, object-identity, or lighting relation, rather than merely continuing the misleading first-frame interpretation.

1 INTRODUCTION

Recent image-to-video generation models have demonstrated strong capability in generating visually realistic videos with coherent motion and physical dynamics from a single input image (Blattmann et al., 2023; Xing et al., 2024). However, visual realism (Huang et al., 2024) does not necessarily imply physically consistent scene evolution (Bansal et al., 2025; Li et al., 2025). This limitation becomes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

especially important under real-world 3D visual illusions (Todorović, 2020; Yao et al., 2025b), where the first-frame appearance can conflict with the observable spatial (Cutting & Vishton, 1995), optical (Fleming, 2014), projection (Raskar et al., 2001; Bimber & Raskar, 2005), surface (Baltrušaitis, 1977), object-identity (Liu et al., 2014), or lighting relation (Kersten et al., 1997; Mamassian et al., 1998). For example, a forced-perspective image may make two objects appear to touch or share the same scale (Kilpatrick & Ittelson, 1953), although the scene provides observable depth cues indicating that the two objects are separated in depth (Cutting & Vishton, 1995). Such illusions are cases where appearance and an observable constraint diverge; under camera or object motion, the dynamics should remain compatible with that constraint. This makes 3D visual illusions a natural stress test for whether image-to-video models can go beyond appearance-level continuation, as illustrated in Figure 1.

Existing video generation benchmarks have made important progress in evaluating visual quality (Huang et al., 2024), temporal consistency (Huang et al., 2024), motion smoothness (Huang et al., 2024), prompt following (Sun et al., 2025), physical commonsense (Bansal et al., 2025), object interactions (Sun et al., 2025), and world-modeling capabilities (Li et al., 2025). However, most benchmarks focus on ordinary scenes or clearly specified events, where the first-frame appearance does not itself conflict with an observable constraint (Huang et al., 2024; Sun et al., 2025; Bansal et al., 2025; Li et al., 2025). As a result, they do not systematically test whether image-to-video models maintain physically consistent scene evolution when the input image is a 3D visual illusion (Huang et al., 2024; Hou et al., 2026). This represents a benchmark gap: current datasets and evaluation protocols do not directly cover failures in this illusion setting.

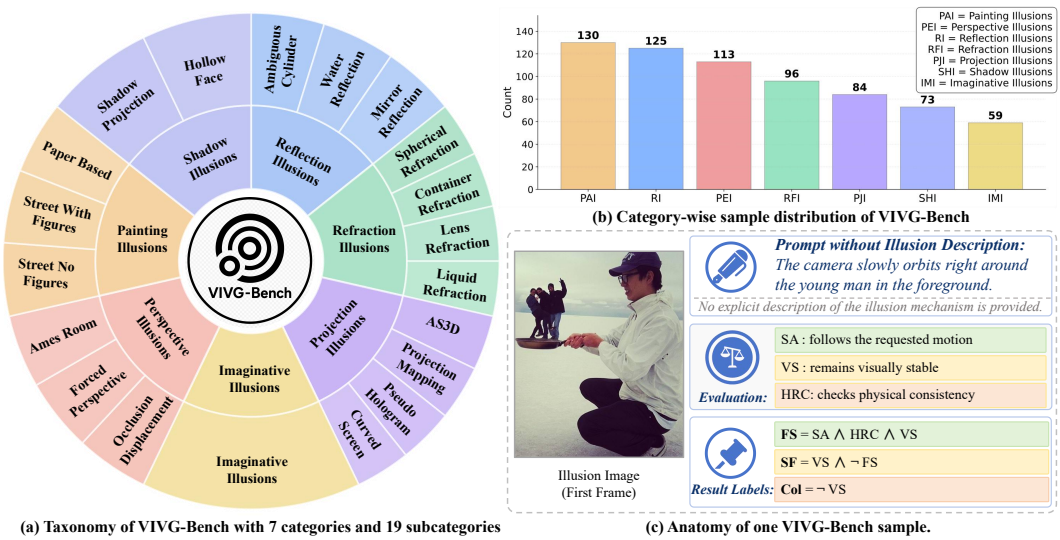
To systematically evaluate this underexplored setting, we introduce VIVG-Bench (**V**isual **I**llusion **V**ideo **G**eneration Benchmark), a benchmark for evaluating image-to-video generation models on real-world 3D visual illusions. VIVG-Bench contains 680 curated image-to-video prompts spanning 7 major categories and 19 fine-grained subtypes, with Imaginative treated as an atomic category without further subtypes. Each sample uses a real-world illusion image as the first frame and pairs it with a targeted camera motion, object motion, lighting change, or local interaction. To avoid information leakage, we use prompts without illusion descriptions: each prompt specifies only the visible entities and intended motion or interaction, without describing the illusion mechanism or revealing the expected physical outcome.

We evaluate state-of-the-art image-to-video generation models, including both closed-source and open-source systems (Google Cloud, 2026; Wang et al., 2025), using a human-centered evaluation protocol. Each generated video is judged along three axes: semantic/action adherence (SA), hidden-relation consistency (HRC), and video stability (VS) (Huang et al., 2024; Liu et al., 2024). Here, video stability means that the generated video remains visually and structurally coherent without severe object deformation, severe appearance or structural drift, background jumps, or global collapse (Huang et al., 2024). Based on SA, HRC, and VS, we derive three outcome labels: Full Success (FS), Stable Failure (SF), and Collapse (Col), as defined in Section 4. Full Success requires the generated video to follow the requested motion, remain consistent with observable geometric, optical, and structural constraints, and remain visually stable. Stable Failure captures videos that remain visually stable but fail to achieve full success, often because the dynamics contradict an observable constraint or do not provide enough motion to test it. Our results show that even the best-performing model achieves only a full-success rate of 65.01%, as shown in Table 1, revealing that visually stable videos in this setting often remain incompatible with observable constraints.

Our contributions are threefold. First, we introduce VIVG-Bench, a benchmark of 680 labeled image-prompt pairs covering 7 major categories and 19 fine-grained subtypes plus one atomic Imaginative category, where prompts avoid explicit illusion descriptions. Second, we propose a human-centered evaluation protocol based on semantic/action adherence, hidden-relation consistency, and video stability. Third, we systematically evaluate closed-source and open-source image-to-video models, revealing stable but physically inconsistent failures under real-world 3D visual illusions.

2 Related Work

Image-to-Video Generation Models. Recent image-to-video models, including Veo (Google Cloud, 2026), Sora (OpenAI, 2024), Kling (Kuaishou Technology, 2024), Seedance (Gao et al., 2025), Wan (Wang et al., 2025), HunyuanVideo (Kong et al., 2024), CogVideo (Yang et al., 2025),



(a) Taxonomy of VIVG-Bench with 7 categories and 19 subcategories

(b) Category-wise sample distribution of VIVG-Bench

(c) Anatomy of one VIVG-Bench sample.

Figure 2: **Overview of VIVG-Bench.** VIVG-Bench covers real-world 3D visual illusions with category-wise statistics and image-prompt instances using prompts without illusion descriptions. Detailed taxonomy and evaluation definitions are provided in Sections 3.2 and 4.

LTX-Video (HaCohen et al., 2024), and Stable Video Diffusion (Blattmann et al., 2023), produce realistic and temporally coherent videos from a single image. Most evaluations emphasize visual realism, temporal smoothness, and prompt or image alignment (Huang et al., 2024; Liu et al., 2024). Preserving the input frame is often appropriate in ordinary image-to-video generation; when that frame is an illusion, the same behavior can contradict an observable constraint. Whether these models stay consistent with observable physical constraints when the first frame is an illusion remains less studied (Todorović, 2020; Hou et al., 2026). VIVG-Bench evaluates this setting.

Video Generation Benchmarks. Existing benchmarks cover visual quality, temporal consistency, prompt following, physical commonsense, and world modeling (Huang et al., 2024; Liu et al., 2024; Sun et al., 2025; Bansal et al., 2025; Meng et al., 2025; Bansal et al., 2026; Li et al., 2025). General-purpose benchmarks such as VBench and EvalCrafter primarily measure visual quality and temporal consistency, whereas VideoPhy, PhyGenBench, VideoPhy2, and WorldModelBench target physical commonsense, object interactions, or scene-level constraints. Both groups generally use ordinary or visually unambiguous conditions, where the target relation is explicit or directly visible (Sun et al., 2025; Li et al., 2025). They therefore do not test compatibility with observable constraints when the first frame is an illusion. VIVG-Bench instead uses a misleading first frame and a prompt that does not reveal the illusion mechanism or the expected physical outcome.

Visual Illusion Benchmarks. Static illusion benchmarks, including HallusionBench, IllusionVQA, Illusory VQA, IllusionBench+, and VIA-Bench, test recognition or question answering on misleading images (Guan et al., 2024; Shahgir et al., 2024; Rostamkhani et al., 2025; Zhang et al., 2025; Hou et al., 2026). They do not test whether a video model maintains physically consistent scene evolution when an illusion image is the first frame. A correct static answer does not determine whether later motion stays compatible with the observable constraint. VIVG-Bench evaluates that generation setting under prompts without illusion descriptions.

3 VIVG-Bench

VIVG-Bench evaluates whether image-to-video generation models produce scene evolution consistent with observable geometric, optical, and structural constraints when the first frame is an illusion. Unlike general image-to-video benchmarks, where the first frame typically provides an unambiguous scene condition, VIVG-Bench uses real-world 3D visual illusions as first-frame inputs. In these cases, the first-frame appearance can conflict with an observable spatial, optical, projection, surface, object-

identity, or lighting relation. A successful video should follow the requested motion or interaction without contradicting that observable constraint.

Each benchmark instance consists of an illusion image and a prompt without illusion description. The prompt specifies only an operational camera motion, object motion, lighting change, or local interaction, without describing the illusion mechanism or revealing the expected physical outcome. Figure 2 provides an overview of the taxonomy, sample distribution, and image-prompt format.

Generated videos are evaluated by semantic/action adherence (SA), hidden-relation consistency (HRC), and video stability (VS), which are combined into Full Success (FS), Stable Failure (SF), and Collapse (Col) labels. Formal definitions are provided in Section 4.

3.1 Task Formulation

Given an input illusion image I_i and a prompt without illusion description p_i, an image-to-video generation model M generates:

$$V_i = M(I_i, p_i).$$

Here, I_i is a real-world 3D visual illusion image used as the first frame, and p_i specifies the intended motion or interaction without describing the illusion mechanism or revealing the expected physical outcome. The goal of VIVG-Bench is relation-consistent generation: V_i should follow the requested motion or interaction, remain consistent with observable geometric, optical, and structural constraints, and remain visually stable. These requirements correspond to SA, HRC, and VS, respectively, which are formally defined in Section 4. The benchmark does not require a unique 3D reconstruction or a particular human perceptual interpretation; multiple compatible explanations can pass. Unlike ordinary image-to-video generation (Zhang et al., 2023; Xing et al., 2024), illusions make appearance diverge from an observable constraint. Preserving the first-frame appearance fails only when the motion contradicts an observable constraint, such as depth separation, surface contact, object continuity, optics, or lighting. A common failure mode is therefore stable failure, where the video remains temporally coherent but does not achieve full success.

3.2 Taxonomy of Real-World 3D Visual Illusions

VIVG-Bench focuses on real-world 3D visual illusions rather than purely 2D optical illusions (Eagleman, 2001; Todorović, 2020). We organize samples by the observable physical relation or structural constraint that must be preserved or updated during video generation, instead of by perceptual effect alone (Gregory, 1997; Hamburger, 2016). This task-oriented taxonomy contains 7 major categories and 19 fine-grained subtypes; Imaginative is an atomic category without further subtypes. The seven categories are reflection (Fleming, 2014), refraction (Fleming, 2014), projection (Raskar et al., 2001; Bimber & Raskar, 2005), imaginative (Liu et al., 2014), painting (Baltrušaitis, 1977), perspective (Kilpatrick & Ittelson, 1953; Cutting & Vishton, 1995), and shadow illusions (Kersten et al., 1997; Mamassian et al., 1998).

The taxonomy is directly tied to HRC evaluation: different categories require different checks, including reflection consistency, refraction consistency, projection-surface dependency, preservation of observable object identity for imaginative illusions, flat-surface preservation for painting illusions, depth/scale/occlusion consistency for perspective illusions, and light-object-surface consistency for shadow illusions. The complete taxonomy, fine-grained sample counts, and category definitions are provided in Appendix A. Representative examples of all 19 fine-grained subtypes and the atomic Imaginative category are shown in Figure 6.

3.3 Benchmark Construction

We construct VIVG-Bench through a taxonomy-guided and human-in-the-loop pipeline, as summarized in Figure 3. Candidate images are collected from publicly available web sources, educational pages, illusion galleries, and existing visual-illusion benchmarks or collections, including HallusionBench (Guan et al., 2024), IllusionVQA (Shahgir et al., 2024), IllusionBench/IllusionBench+ (Zhang et al., 2025), and VIA-Bench (Hou et al., 2026). We use these resources only as candidate image sources and do not inherit their static-image task formats.

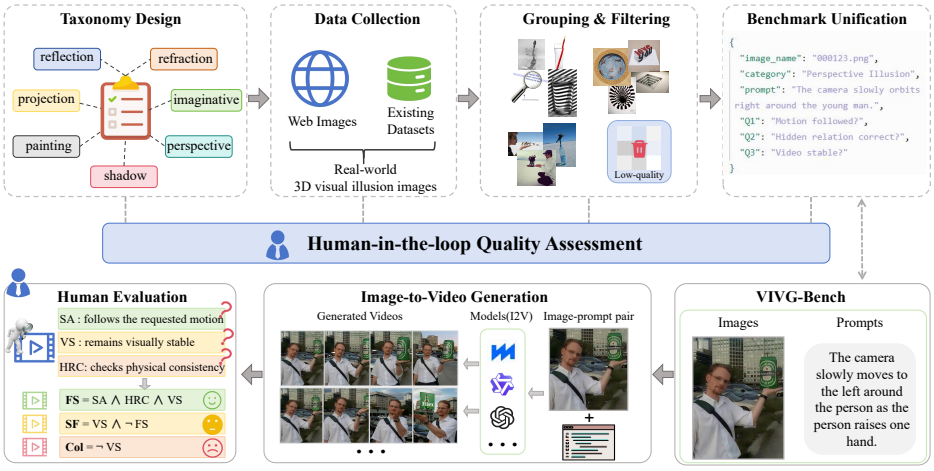


Figure 3: **Construction and evaluation pipeline of VIVG-Bench.** We construct VIVG-Bench through taxonomy-guided collection, filtering, and image-prompt unification, then evaluate generated videos with human SA/HRC/VS annotations. Details are provided in Sections 3.3 and 4.

We manually filter candidates according to two criteria. First, each image must depict a real-world 3D visual illusion rather than a purely 2D geometric, color, or static-motion illusion (Eagleman, 2001; Todorović, 2020). Second, the illusion must be dynamically evaluable: there should exist a camera motion, object interaction, lighting change, or local physical event that can test the relevant physical relation during generation. Images are removed when observable scene evidence is insufficient to evaluate the relevant physical relation under the specified intervention, as well as when they are weakly illusory, duplicated, low-quality, or difficult to evaluate through motion.

For each retained image, we assign a major category and fine-grained type, identify the relevant observable physical relation or structural constraint, and design a prompt without illusion description. Images were screened to center on one dominant 3D illusion whenever possible. Samples with additional visual effects were retained but assigned one primary illusion type, and the prompt targets that dominant physical relation. A systematic evaluation of images with multiple coexisting illusions would require multi-label annotations and is left for future work. Each prompt is checked to ensure that it provides a meaningful dynamic condition without describing the illusion mechanism or revealing the expected physical outcome. The final benchmark is unified as image-prompt pairs, where each image serves as the first frame and each prompt specifies the intended dynamic transformation. Detailed category definitions are provided in Appendix A.

This focus on real-world 3D illusions is also motivated by recent findings in 3D visual illusion depth estimation. Yao et al. (2025a) show that modern monocular, binocular, and multi-view depth estimation methods can be seriously misled by 3D visual illusions. VIVG-Bench tests the same illusion setting in video: whether generated dynamics remain compatible with observable constraints.

3.4 Prompts without Illusion Descriptions and Dataset Statistics

A central design principle of VIVG-Bench is the prompt without illusion description. Rather than describing the scene or explaining the illusion, the prompt specifies only an operational motion or interaction that makes the relevant physical relation testable. For example, it may ask the camera to orbit around a forced-perspective scene, an object to roll across a painted surface, or a light source to move in a shadow illusion. Among the 680 prompts, 328 primarily specify camera motion, 315 primarily specify object or person motion, and 37 primarily specify lighting changes, including moving the light source. Prompts were screened to center on one primary operation; multi-operation prompts, prompt sensitivity, and coexisting illusions remain outside the current scope.

Prompts without illusion descriptions avoid explicit descriptions such as “this is a flat painting,” “this is a forced-perspective illusion,” or “the apparent object is not real.” They also avoid instructing the model to produce the expected physical outcome directly. Thus, the prompt may induce motion, but

Table 1: **Main results on VIVG-Bench.** We report SA, HRC, VS, FS, SF, and Col as percentages over returned clips; safety-filter blocks are excluded from the denominators. All metric definitions are in Section 4. Higher is better for SA, HRC, VS, and FS, while lower is better for SF and Col. Dark green and light green indicate the best and second-best results, respectively.

Model	SA ↑	HRC ↑	VS ↑	FS ↑	SF ↓	Col ↓
Closed-source Models						
Grok Imagine (xAI, 2026)	98.56	65.81	96.47	65.01	31.46	3.53
Veo 3.1 (Google Cloud, 2026)	97.10	66.34	90.50	62.32	28.18	9.50
Veo 3.1 Fast (Google Cloud, 2026)	97.36	61.59	89.89	58.63	31.26	10.11
Seedance 1.0 Pro Fast (Seed Vision Team, 2025a)	90.16	54.06	83.75	52.19	31.56	16.25
Seedance 1.5 Pro (Seed Vision Team, 2025b)	97.52	47.29	94.26	45.12	49.15	5.74
Wan2.6-I2V-Flash (Alibaba Tongyi Lab, 2026)	90.21	47.59	85.69	44.13	41.57	14.31
Open-source Models						
Wan2.2-I2V-Plus (Wan Team et al., 2025)	79.13	58.71	70.57	55.86	14.71	29.43
CogVideoX-3 (Yang et al., 2024)	88.61	55.18	96.15	50.89	45.27	3.85

it does not provide the illusion category, the cause of the illusion, or the expected physical outcome. Representative prompt examples are provided in Appendix B.

VIVG-Bench contains 680 high-quality image-prompt pairs across 7 major categories and 19 fine-grained subtypes plus one atomic Imaginative category. All main tables use the original frozen 680-sample manifest and its adjudicated labels. Provider requests blocked by safety filters produced no clip and are excluded from the rate denominators rather than imputed; availability records are in the release. Each instance includes an image identifier, the input illusion image, the major category, the fine-grained type, and the prompt without illusion description. Together, these samples provide a diagnostic testbed for evaluating whether image-to-video models can go beyond appearance-level continuation and remain consistent with the observable physical constraints when the first frame is an illusion.

4 Evaluation Protocol

As introduced in Section 3, VIVG-Bench evaluates each generated video using three binary criteria: semantic/action adherence (SA), hidden-relation consistency (HRC), and video stability (VS). These criteria ask whether the video follows the requested motion, remains consistent with the observable physical relation or structural constraint supported by the input scene, and remains visually stable. They are combined into three outcome labels: Full Success (FS), Stable Failure (SF), and Collapse (Col). This section formally defines the criteria and describes the human annotation protocol, with additional annotation details, agreement statistics, and automatic-judge analysis provided in Appendices D, D.1, and D.2.

4.1 Evaluation Criteria

For each generated video, annotators are given the first-frame illusion image, the prompt without illusion description, and the generated video. They evaluate the video using three binary criteria: semantic/action adherence (SA), hidden-relation consistency (HRC), and video stability (VS).

Semantic/action adherence (SA). SA measures whether the generated video follows the requested motion or interaction specified by the prompt without illusion description. This criterion checks prompt-level execution, but does not by itself indicate whether the observable physical relation is handled consistently.

Hidden-relation consistency (HRC). HRC measures whether the generated video remains consistent with the physical relation or structural constraint supported by the observable evidence in the input scene. Unlike generic physical plausibility, HRC is category-aware: different illusion categories require different checks, including reflection consistency, refraction consistency, projection-surface de-

Table 2: **Category-wise results on VIVG-Bench.** We report HRC and FS as percentages for each illusion category; provider-blocked requests are excluded from these denominators. Definitions are provided in Section 4. Category abbreviations follow the taxonomy in Appendix A, and rank is based on average FS. Dark green and light green indicate the best and second-best results, respectively.

Model	Avg.		Rank	RI		RFI		PJI		IMI		PAI		PEI		SHI	
	HRC	FS		HRC	FS	HRC	FS	HRC	FS	HRC	FS	HRC	FS	HRC	FS	HRC	FS
Closed-source Models																	
Grok Imagine	65.81	65.01	1	78.51	78.51	69.15	69.15	46.34	46.34	93.10	93.10	74.19	73.39	57.69	57.69	32.81	32.81
Veo 3.1	66.34	62.32	2	70.54	64.29	75.56	68.89	48.72	47.44	96.55	94.83	77.19	76.32	60.19	54.37	31.82	27.27
Veo 3.1 Fast	61.59	58.63	3	63.87	55.46	68.82	67.74	53.75	53.75	81.36	81.36	64.10	64.10	54.81	47.12	46.48	46.48
Seedance 1.0 Pro Fast	54.06	52.19	5	60.00	56.52	81.72	79.57	52.44	48.78	75.93	72.22	27.20	27.20	55.24	55.24	37.88	36.36
Seedance 1.5 Pro	47.29	45.12	7	49.14	46.55	65.96	63.83	43.21	41.98	52.73	50.91	25.20	23.62	63.21	60.38	34.85	31.82
Wan2.6-I2V-Flash	47.59	44.13	8	50.85	43.22	46.32	43.16	36.90	32.14	66.07	60.71	44.19	42.64	56.25	56.25	34.29	31.43
Open-source Models																	
Wan2.2-I2V-Plus	58.71	55.86	4	60.17	51.69	81.05	78.95	42.86	41.67	73.68	70.18	38.76	36.43	62.83	62.83	62.86	61.43
CogVideoX-3	55.18	50.89	6	67.20	60.80	72.92	62.50	41.67	39.29	94.83	86.21	40.16	37.80	54.39	53.51	22.22	22.22

pendency, preservation of observable object identity for imaginative illusions, flat-surface preservation for painting illusions, depth/scale/occlusion consistency for perspective illusions, and light-object-surface consistency for shadow illusions. HRC does not ask whether a model shares a particular human perceptual interpretation of the initial image. An HRC failure requires a generated scene evolution that contradicts the observable physical constraints of the input scene. Full category-aware HRC rubrics used by annotators are provided in Appendix D.

Video stability (VS). VS measures whether the generated video remains visually and structurally stable. A video is considered unstable if it exhibits severe object deformation, severe appearance or structural drift, structural collapse, abrupt background jumps, or global scene instability. Local deformation that leaves object identity, structure, and scene continuity intact still receives $\mathrm{VS} = 1$; only severe or persistent collapse sets VS to 0. Semantic identity errors that remain temporally and structurally coherent are evaluated by HRC rather than VS; they do not constitute Collapse by themselves. VS separates failures of physical-relation consistency from low-level generation failures: a video may violate an observable physical constraint while still remaining visually stable, or it may collapse before that constraint can be meaningfully evaluated.

Based on these three criteria, we derive the final outcome labels used throughout the experiments:

$$\mathrm{FS} = \mathrm{SA} \wedge \mathrm{HRC} \wedge \mathrm{VS}, \quad \mathrm{Col} = \neg\mathrm{VS}, \quad \mathrm{SF} = \mathrm{VS} \wedge \neg\mathrm{FS}.$$

Full Success (FS) requires the video to follow the requested motion or interaction, remain consistent with the relevant observable physical constraints, and remain visually stable. **Collapse (Col)** denotes visually or structurally unstable generations. **Stable Failure (SF)** is an outcome label rather than a single failure mechanism. A visually stable video receives this label whenever it does not satisfy FS, which may result from physical-relation inconsistency, failure to execute the requested action, or another unmet FS criterion.

4.2 Human Annotation

We use human annotation as the primary evaluation protocol. For each generated video, annotators are shown the first-frame illusion image, the prompt without illusion description, and the generated video, and independently judge SA, HRC, and VS. The final FS/SF/Col label is then derived from these three binary criteria.

Each generated video is annotated by two annotators, with disagreements resolved by a third annotator through adjudication. Annotators use category-aware HRC rubrics to judge constraint violations, not a lay preference for one perceptual reading. All main experimental results are computed from adjudicated human labels. Detailed annotation instructions, annotator recruitment, blind evaluation procedure, and category-aware HRC rubrics are provided in Appendix D.

We also explored automatic judging with off-the-shelf MLLM judges (Chen et al., 2024; Alibaba Cloud, 2026), but found limited agreement with human annotations. Therefore, automatic judging is used only as an auxiliary analysis, with details provided in Appendix D.2.

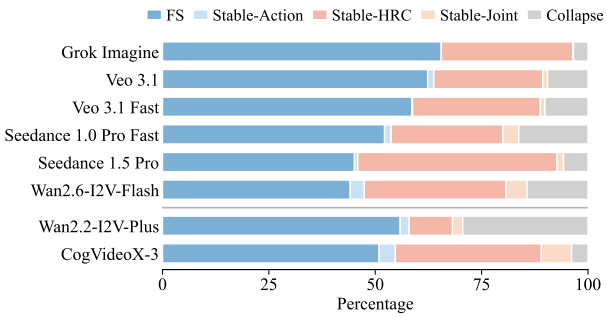


Figure 4: Overall outcome profiles of the evaluated image-to-video models on VIVG-Bench. Bars show mutually exclusive Full Success, Stable-Action, Stable-HRC, Stable-Joint, and Collapse profiles. These profiles highlight distinct failure modes that are obscured by aggregate video-quality scores.

Agreement from the actual pre-adjudication labels is high: the final FS/SF/Col label reaches 90.4% agreement with a Cohen's κ of 0.838, and HRC reaches 92.5% agreement with a Cohen's κ of 0.846. Detailed overall and category-wise agreement statistics are provided in Appendix D.1.

5 Experiments

5.1 Experimental Setup

We evaluate six closed-source and two open-source image-to-video models under the same setting: a real-world illusion image as first frame and a prompt without illusion description. Each model uses its official shortest or recommended duration; settings and valid-clip handling are in Appendix C. All models use the same frozen 680 image-prompt instances. Metrics are computed over returned clips with complete labels; safety-filter blocks are excluded rather than counted as Collapse. We report SA, HRC, VS, FS, SF, and Col, with higher values preferred for SA/HRC/VS/FS and lower values for SF/Col.

5.2 Main Results

Table 1 reports overall performance, while Figure 4 decomposes outcomes into mutually exclusive profiles. Collapse (VS $= 0$) takes priority; otherwise Stable Failure is separated by whether SA, HRC, or both are missed (Section 4.1). Stable-Action misses only SA, Stable-HRC misses only HRC, and Stable-Joint misses both; all three keep VS $= 1$. Stable-HRC is the dominant non-success pattern for Seedance 1.5 Pro, whereas Wan2.2-I2V-Plus has more Collapse; Stable-Action and Stable-Joint remain rare.

A high VS does not guarantee physical consistency: CogVideoX-3 has high VS but lower FS, showing stable HRC failures. Grok Imagine achieves the highest FS and Veo 3.1 the highest HRC, while a low SF can coexist with high Collapse (e.g., Wan2.2-I2V-Plus). Thus, visual realism and stability remain distinct from relation-consistent scene evolution. A lower SF rate can reflect either more Full Success or more Collapse. Seedance 1.5 Pro illustrates the first pattern: most of its non-success is Stable-HRC rather than Collapse. Wan2.2-I2V-Plus illustrates the second, with Collapse at 29.43% and a much smaller Stable-HRC share.

5.3 Category-wise Analysis

Table 2 reports category-wise HRC and FS, revealing substantial variation rather than a single generic capability. We emphasize HRC because it directly tests the category-specific observable constraint, while FS additionally requires action adherence and stability. Imaginative performance is relatively strong, while Projection and Shadow remain challenging; Perspective, Reflection, and Refraction expose distinct depth, occlusion, and optical failures. These observations are descriptive rather than causal: the evaluated systems do not expose their training data or internal mechanisms. The output-level differences are consistent with object-identity changes, projection-surface violations, depth or occlusion errors, and light-object-surface inconsistencies. The cases in Figure 5 further show that these are distinct, often visually stable failure patterns.

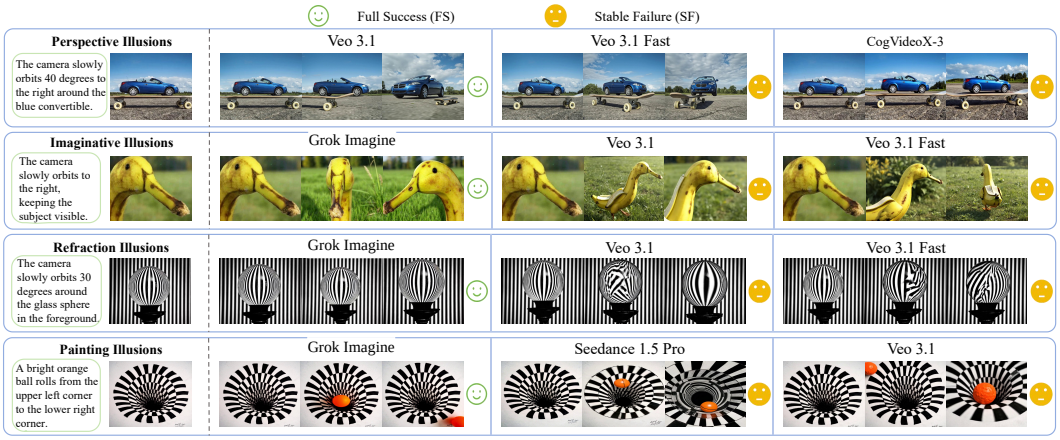


Figure 5: **Qualitative comparison on VIVG-Bench.** Representative FS and SF examples across illusion categories under prompts without illusion descriptions, with outcomes annotated directly in the figure.

5.4 Qualitative Analysis

Figure 5 contrasts Full Success and Stable Failure on perspective, refraction, imaginative, and painting examples. The Stable Failure clips stay temporally coherent and often carry out the requested motion, yet they contradict an observable constraint: forced-perspective depth is left unchanged under the orbit, refraction stays nearly fixed as the viewpoint moves, a resemblance is turned into a real object, or a painted structure is treated as solid geometry. Full Success uses the same prompt to update that relation: the orbit reveals the forced-perspective separation, the refracted pattern moves with the viewpoint, the object keeps its visible material and structure, and the ball remains on the flat painted surface. On each row, one model can reach Full Success while another produces Stable Failure, so the failure is not explained by an impossible prompt.

6 Discussion and Limitations

Prompt-side analyses provide two controls on Grok Imagine (Appendix E). Removing the motion prompt lowers a diagnostic score from the benchmark HRC of 65.81% to 46.51% while VS stays at 93.18%; SA, FS, and SF are undefined without a target action, and this score is not the main HRC label. On a re-selected paired subset, stating that a visual illusion is present raises HRC from 65.81% to 69.50% and FS from 65.65% to 68.54%, and raises Collapse from 3.53% to 5.14% (Appendix E.2). The cue yields only a modest gain. A lower Stable Failure rate can mean more Full Success or more Collapse. Agreement is high overall ($\kappa = 0.838$). The benchmark uses short clips with one dominant intervention and does not include matched ordinary scenes. These results indicate that prompt-side cues can partially improve hidden-relation consistency, but do not eliminate physically inconsistent generations. More broadly, the findings suggest that explicit illusion awareness remains insufficient to fully resolve the gap between visual stability and constraint-consistent motion.

7 Conclusion

VIVG-Bench pairs misleading illusion images with prompts without illusion descriptions. Human evaluation shows that stable but physically inconsistent videos remain common. Separating action adherence, hidden-relation consistency, and visual stability makes these failures measurable rather than a single notion of video quality. These illusions are cases where appearance and an observable constraint diverge. The benchmark does not require a unique 3D reconstruction: a video passes when its motion remains compatible with geometric, optical, and structural constraints observable from the input, and multiple compatible explanations can pass. Under this criterion, even the best model reaches only a 65.01% full-success rate, because many visually stable videos still violate those constraints. The results describe this illusion setting; they do not show that illusions cause additional failures on ordinary scenes.

AI use statement. Generative AI assisted with language editing and data processing; the authors reviewed all assisted work and take responsibility for the paper's content and artifacts.

Ethics statement. This work evaluates image-to-video models on publicly available visual-illusion images and the clips generated from them. Human labels come from three compensated annotators; the annotation task does not collect sensitive personal information. The study does not deploy a system that acts on people or the physical environment.

Reproducibility statement. The paper and appendices document benchmark construction, prompts, evaluation, and annotation. The anonymized repository (https://huggingface.co/datasets/VIVG-Bench/VIVG-Bench) provides the frozen 680-sample manifest, data, code, labels, instructions, and representative clips, but not every generation. The repository and these artifacts must be accessible to reviewers at submission time so that the reported tables can be checked against the same release.

REFERENCES

Alibaba Cloud. Supported Models and Capabilities Overview: Qwen3-VL-Flash Series, 2026. URL https://www.alibabacloud.com/help/en/model-studio/models. Accessed: 2026-05-07.

Alibaba Tongyi Lab. Wan2.6 image-to-video flash. https://replicate.com/wan-video/wan2.6-i2v-flash/readme, 2026. Accessed: 2026-05-07.

Jurgis Baltrušaitis. *Anamorphic Art*. Harry N. Abrams, New York, 1977. ISBN 0-8109-0662-7. Translated by W. J. Strachan.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9D2QvO1uWj.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HA8KSQW7SO.

Oliver Bimber and Ramesh Raskar. *Spatial Augmented Reality: Merging Real and Virtual Worlds*. A K Peters, Wellesley, MA, 2005. ISBN 9781568812304.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *International Conference on Machine Learning*, 2024.

James E. Cutting and Peter M. Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In William Epstein and Sheena Rogers (eds.), *Perception of Space and Motion*, pp. 69–117. Academic Press, San Diego, 1995. doi: 10.1016/B978-0-12-240530-3.50005-5.

David M. Eagleman. Visual illusions and neurobiology. *Nature Reviews Neuroscience*, 2(12): 920–926, 2001. doi: 10.1038/35104092.

Roland W. Fleming. Visual perception of materials and their properties. *Vision Research*, 94:62–75, 2014. doi: 10.1016/j.visres.2013.11.004.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025. doi: 10.48550/arXiv.2506.09113.

Google Cloud. Veo 3.1. https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/veo/3-1-generate, 2026. Accessed: 2026-05-07.

Richard L. Gregory. Visual illusions classified. *Trends in Cognitive Sciences*, 1(5):190–194, 1997. doi: 10.1016/S1364-6613(97)01060-7.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. doi: 10.48550/arXiv.2501.00103.

Kai Hamburger. Visual illusions based on processes: New classification system needed. *Perception*, 45(5):588–595, 2016. doi: 10.1177/0301006616629038.

Wenjin Hou, Wei Liu, Han Hu, Xiaoxiao Sun, Serena Yeung-Levy, and Hehe Fan. Seeing is believing? a benchmark for multimodal large language models on visual illusions and anomalies. *arXiv preprint arXiv:2602.01816*, 2026. doi: 10.48550/arXiv.2602.01816.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Daniel Kersten, Pascal Mamassian, and David C. Knill. Moving cast shadows induce apparent motion in depth. *Perception*, 26(2):171–192, 1997. doi: 10.1068/p260171.

Franklin P. Kilpatrick and William H. Ittelson. The size-distance invariance hypothesis. *Psychological Review*, 60(4):223–231, 1953. doi: 10.1037/h0060882.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Kuaishou Technology. Kuaishou unveils proprietary video generation model "kling"; testing now available, 2024. URL https://www.prnewswire.com/news-releases/kuaishou-unveils-proprietary-video-generation-model-kling-testing-now-available-302168757.html. Accessed: 2026-05-07.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=a3hafrDzuA.

Jiangang Liu, Jun Li, Lu Feng, Ling Li, Jie Tian, and Kang Lee. Seeing jesus in toast: Neural and behavioral correlates of face pareidolia. *Cortex*, 53:60–77, 2014. doi: 10.1016/j.cortex.2014.01.013.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Pascal Mamassian, David C. Knill, and Daniel Kersten. The perception of cast shadows. *Trends in Cognitive Sciences*, 2(8):288–295, 1998. doi: 10.1016/S1364-6613(98)01204-2.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43781–43806. PMLR, 2025. URL https://proceedings.mlr.press/v267/meng25c.html.

OpenAI. Video generation models as world simulators, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/. Accessed: 2026-05-07.

Ramesh Raskar, Greg Welch, Kok-Lim Low, and Deepak Bandyopadhyay. Shader lamps: Animating real objects with image-based illumination. In *Rendering Techniques 2001: Proceedings of the 12th Eurographics Workshop on Rendering*, pp. 89–102. Springer, 2001. doi: 10.2312/EGWR/EGWR01/089-101.

Mohammadmostafa Rostamkhani, Baktash Ansari, Hoorieh Sabzevari, Farzan Rahmani, and Sauleh Eetemadi. Illusory vqa: Benchmarking and enhancing multimodal models on visual illusions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2995–3004, 2025. doi: 10.1109/CVPRW67362.2025.00282.

Seed Team. Seed2.0. Doubao-Seed-2.0-lite model page, 2026. URL https://seed.bytedance.com/en/seed2. Accessed: 2026-05-07.

Seed Vision Team. Seedance 1.0: Exploring the boundaries of video generation models. https://seed.bytedance.com/public_papers/seedance-1-0-exploring-the-boundaries-of-video-generation-models, 2025a. Accessed: 2026-05-07.

Seed Vision Team. Seedance 1.5 pro: A native audio-visual joint generation foundation model. https://seed.bytedance.com/public_papers/seedance-1-5-pro-a-native-audio-visual-joint-generation-foundation-model, 2025b. Accessed: 2026-05-07.

Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. Illusionvqa: A challenging optical illusion dataset for vision language models. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=7ysaJGs7zY.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Dejan Todorović. What are visual illusions? *Perception*, 49(11):1128–1199, 2020. doi: 10.1177/0301006620962279.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

xAI. Grok imagine video. https://replicate.com/xai/grok-imagine-video, 2026. Accessed: 2026-05-07.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. In *Proceedings of the European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-72952-2_23.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Chengtang Yao, Zhidan Liu, Jiaxi Zeng, Lidong Yu, Yuwei Wu, and Yunde Jia. 3d visual illusion depth estimation. In *Advances in Neural Information Processing Systems*, 2025a.

Chengtang Yao, Zhidan Liu, Jiaxi Zeng, Lidong Yu, Yuwei Wu, and Yunde Jia. 3D visual illusion depth estimation. *arXiv preprint arXiv:2505.13061*, 2025b. doi: 10.48550/arXiv.2505.13061.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. doi: 10.48550/arXiv.2311.04145.

Yiming Zhang, Zicheng Zhang, Xinyi Wei, Xiaohong Liu, Guangtao Zhai, and Xiongkuo Min. Illusionbench+: A large-scale and comprehensive benchmark for visual illusion understanding in vision-language models. *arXiv preprint arXiv:2501.00848*, 2025. doi: 10.48550/arXiv.2501.00848.

APPENDIX

A TAXONOMY DETAILS

Table 3 provides the complete taxonomy of VIVG-Bench. The benchmark contains 7 major categories and 19 fine-grained subtypes; Imaginative is an atomic category without further subtypes. We use RI, RFI, PJI, IMI, PAI, PEI, and SHI to denote Reflection, Refraction, Projection, Imaginative, Painting, Perspective, and Shadow Illusions, respectively. Imaginative Illusions do not contain further subcategories and are treated as a single fine-grained type.

The seven major categories are defined by the observable physical relation or structural constraint that must be preserved or updated during video generation. **Reflection illusions** involve mirror or water reflections whose reflected content should change consistently under viewpoint or object motion (Fleming, 2014). **Refraction illusions** involve transparent media such as glass, water, lenses, or containers, where optical distortion should remain continuous and physically consistent (Fleming, 2014). **Projection illusions** involve screen-based, projection-mapped, curved-screen, or hologram-like effects, where apparent content should remain constrained by the display or projection medium (Raskar et al., 2001; Bimber & Raskar, 2005). **Imaginative illusions** involve real-world objects or scenes that resemble another semantic entity, where the generated video should preserve the physical identity supported by the observable object, material, and structure, rather than turning the resemblance into a real object (Liu et al., 2014). **Painting illusions** involve flat drawings or surface paintings that imitate 3D structures, where the generated video should preserve the flat-surface nature of the artwork (Baltrušaitis, 1977). **Perspective illusions** involve forced-perspective, Ames-room, or occlusion-displacement effects, where motion should not introduce depth, scale, separation, or occlusion that contradicts the observable evidence (Kilpatrick & Ittelson, 1953; Cutting & Vishton, 1995). **Shadow illusions** involve ambiguous or misleading shadows, where the generated video should preserve consistency among the light source, object geometry, and projection surface (Kersten et al., 1997; Mamassian et al., 1998).

Table 3: **Complete taxonomy of VIVG-Bench with fine-grained sample counts.** VIVG-Bench contains 7 major categories and 19 fine-grained subtypes plus one atomic Imaginative category ($n = 680$).

Major Category	Fine-grained Type	n
Reflection Illusions	Ambiguous Cylinder	24
	Water Reflection	32
	Mirror Reflection	69
Refraction Illusions	Spherical Refraction	18
	Container Refraction	33
	Lens Refraction	21
	Liquid Refraction	24
Projection Illusions	Curved Screen	32
	Pseudo Hologram	28
	Projection Mapping	12
	Autostereoscopic 3D	12
Imaginative Illusions	Imaginative Illusions	59
Painting Illusions	Paper Based	46
	Street With Figures	68
	Street No Figures	16
Perspective Illusions	Ames Room	6
	Forced Perspective	71
	Occlusion Displacement	36
Shadow Illusions	Hollow Face	9
	Shadow Projection	64

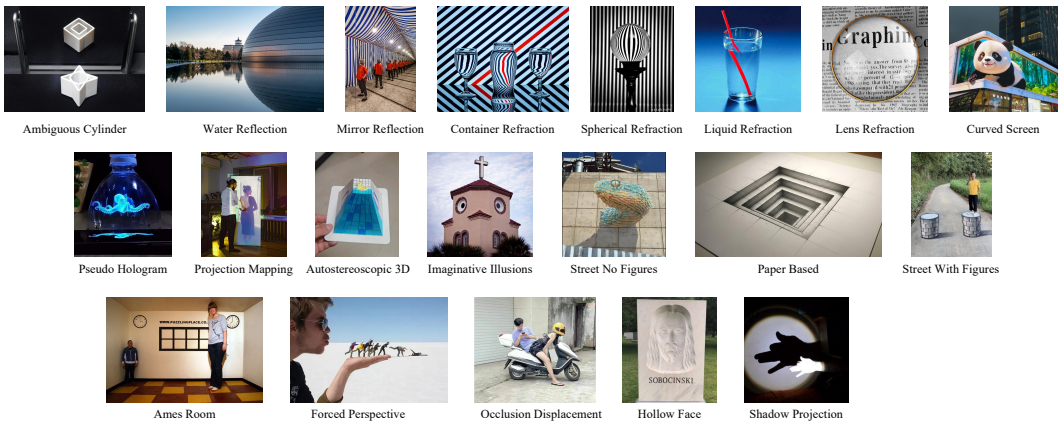


Figure 6: **Representative examples of the fine-grained illusion types in VIVG-Bench.** The figure shows one representative example for each of the 19 fine-grained subtypes and the atomic Imaginative category across the 7 major illusion categories.

Table 4: **Representative prompts without illusion descriptions.** Each prompt induces motion or interaction without revealing the illusion mechanism or the expected physical outcome.

Category	Example Prompt without Illusion Description
Reflection	The camera slowly moves sideways while keeping the reflective surface visible.
Refraction	The camera slowly orbits around the transparent object in the foreground.
Projection	The camera slowly moves to the right while keeping the displayed content visible.
Imaginative	The camera slowly orbits around the object while keeping it centered in view.
Painting	A small ball rolls from the upper-left area to the lower-right area across the visible surface.
Perspective	The camera slowly orbits around the foreground subject.
Shadow	The light source slowly moves from left to right while the scene remains visible.

These fine-grained counts sum to the seven category totals in Figure 2(b): Painting 130, Reflection 125, Perspective 113, Refraction 96, Projection 84, Shadow 73, and Imaginative 59.

B Prompt Examples without Illusion Descriptions

Table 4 provides representative prompts without illusion descriptions for each major illusion category. These prompts specify only an operational motion or interaction and do not describe the illusion mechanism, illusion category, or expected physical outcome.

C Generated Video Lengths

We use each model's official shortest or recommended setting and evaluate the generated videos. Table 5 reports the duration, frame rate, and frame count for every evaluated model. Across these settings, the clips last 4–8 seconds and contain 120–192 frames.

D Human Annotation Details and Category-aware Rubrics

Annotator recruitment and independence. Human evaluation uses a pool of three annotators with an average of about three years of relevant experience. For each generated video, two annotators independently provide the labels; the third annotator is used only to adjudicate disagreements. They did not participate in benchmark construction, model generation, prompt design, or result analysis. Annotators are not informed of the expected model outcomes or the hypotheses of the study. All annotators are compensated for their work. The annotation task involves evaluating illusion images,

Table 5: **Generated video lengths.** Each row is one evaluated model. Duration and frame rate follow that model's official shortest or recommended setting, and the frame count is that of the generated video.

Model	Duration (s)	Frame rate (fps)	Frames
Grok Imagine	6	24	144
Veo 3.1	8	24	192
Veo 3.1 Fast	8	24	192
Seedance 1.0 Pro Fast	5	24	120
Seedance 1.5 Pro	5	24	120
Wan2.6-I2V-Flash	4	30	120
Wan2.2-I2V-Plus	5	30	150
CogVideoX-3	5	30	150

prompts without illusion descriptions, and generated videos, and does not collect sensitive personal information from annotators.

Training and calibration. Before annotation, annotators receive instructions explaining the task setting, the definitions of semantic/action adherence (SA), hidden-relation consistency (HRC), video stability (VS), and the derived Full Success (FS), Stable Failure (SF), and Collapse (Col) labels. They also complete a small calibration set to ensure that they understand the evaluation criteria and the category-aware HRC rubrics. Calibration examples are used only for training and are not included in the reported evaluation results.

Blind and randomized evaluation. The annotation interface presents the first-frame illusion image, the prompt without illusion description, and the generated video. Model names, model families, source labels, and expected outcomes are hidden from annotators. The order of samples and generated videos is randomized to reduce ordering bias. Annotators judge only the actual generated video content, rather than the intended ideal outcome implied by the prompt.

Annotation questions. For each generated video, annotators answer three binary questions corresponding to SA, HRC, and VS. The final label (FS/SF/Col) is derived according to Section 4.

- **SA**: Does the video follow the requested motion or interaction?
- **HRC**: Does the video remain consistent with the relevant observable physical relation or structural constraint?
- **VS**: Does the video remain visually and structurally stable?

Adjudication. Each generated video is independently annotated by two annotators. If the two annotators disagree on any of the three binary criteria or on the derived final label, a third annotator reviews the same image-prompt-video instance and resolves the disagreement. All main experimental results reported in the paper are computed from these adjudicated labels.

Category-aware HRC rubrics. Table 6 gives an operational check for every taxonomy entry. The annotator identifies the visible cue named in the table, checks whether the requested operation is attempted and makes that relation testable, and assigns HRC $= 1$ only when the resulting evidence is compatible with the cue. HRC $= 0$ is assigned for a specified contradiction, an absent operation, or an untestable relation after the attempted operation. SA records operation adherence separately, while severe temporal or structural breakdown is marked by VS. Plausible alternative scene explanations pass only when they remain testable and cue-compatible; disputed cases are referred for adjudication rather than forced into a favorable interpretation.

For sample-level auditability, the accompanying release includes bilingual JSONL records with image-specific SA/HRC/VS questions, pass/fail conditions, and uncertainty flags. Each record is keyed by the stable `case_id`: its first numeric component indexes the frozen 680-sample order used for the original taxonomy, while its second component preserves the source image identifier. The remapped files reproduce the original 19 fine-grained subtype counts plus the atomic Imaginative count and

retain a separate manual-review queue for ambiguous cases. These records define the annotation protocol and audit trail; they are not model outputs or replacement labels for the adjudicated metrics.

Table 6: **Category-aware HRC pass/fail standards.** Annotators use the two conditions below after confirming that the requested operation is attempted and the relation is testable.

Taxonomy entry	HRC pass	HRC fail
Ambiguous Cylinder	The contour and mirror or projection boundary remain geometrically compatible; apparent continuity may separate under viewpoint motion.	Separate forms become an unsupported connected object, or the continuation moves independently of the boundary.
Water Reflection	Reflected content changes with the source and water surface while remaining on the reflective surface.	The reflection is a fixed texture, moves independently of its source, or detaches from the water boundary.
Mirror Reflection	The virtual image updates with source and viewpoint motion and stays in the mirror relation.	The image is a fixed texture, becomes an independent object, or crosses the mirror plane without support.
Spherical Refraction	Distortion remains attached to the sphere and changes continuously with viewing geometry.	The optical effect detaches from the sphere or changes incompatibly with the sphere and background.
Container Refraction	Displacement remains bounded by the transparent medium and follows container, liquid, and viewpoint changes.	Displacement appears outside the medium or ignores visible viewing-geometry changes.
Lens Refraction	Magnification or displacement remains within the lens aperture and follows lens position.	The effect leaves the aperture, remains fixed while the lens moves, or changes discontinuously without cause.
Liquid Refraction	Distortion remains associated with the liquid boundary and changes continuously with the medium.	Distortion crosses the boundary without material change or remains rigid despite liquid motion.
Curved Screen	Displayed content stays registered to the curved display surface and follows its curvature.	Content separates as a free-standing object or stays fixed while the display moves.
Pseudo Hologram	The apparent image remains compatible with the visible display, reflector, or projection support.	The image acquires unsupported depth or interacts as a free-standing physical object.
Projection Mapping	Projected content remains registered to the receiving surface and follows its shape.	Content leaves the surface or stays fixed while the receiving surface moves.
Autostereoscopic 3D	Apparent depth remains a display effect bounded by the screen.	The apparent object crosses the screen boundary or interacts with the external scene without support.
Imaginative Illusions	Observable material and structure remain continuous; multiple compatible interpretations pass.	Semantic resemblance is materialized as a different physical entity or visible source structure is contradicted.
Paper Based	Depicted depth remains part of the flat artwork and follows the paper surface.	The drawing becomes solid geometry or interacts as a separate structure without support.
Street With Figures	Real figures remain on the ground plane and painted content remains a surface depiction.	A figure enters painted depth, or paint rises as solid geometry without visible construction.
Street No Figures	Apparent depth remains a flat-surface effect attached to the continuous ground plane.	A painted void or object becomes real geometry despite the visible ground.
Ames Room	Motion remains compatible with visible room construction and scale; compatible alternative constructions pass.	Motion violates visible room geometry or creates an impossible scale relation.
Forced Perspective	Alignment changes consistently with visible depth and occlusion cues.	Motion forces contact, equal depth, or scale relations ruled out by the image.
Occlusion Displacement	Newly exposed parts join the visible parts under one consistent object and occlusion layout.	Revealed parts cannot belong to the same object or require an impossible occlusion change.
Hollow Face	The generated concave/convex interpretation and subsequent shading remain compatible with the initial visible facial surface, occlusion, and lighting cues; any interpretation compatible with those initial cues passes.	Viewpoint or light motion reverses or contradicts the initial surface, occlusion, or lighting cues, or creates geometry inconsistent with them.
Shadow Projection	Shadow position and shape remain consistent with the light, casting object, and receiving surface.	Shadow changes independently, detaches from the surface, or contradicts the visible light direction.

D.1 Human Annotation Agreement

Table 7 reports pre-adjudication agreement from the actual two independent annotators for SA, HRC, VS, and the derived final label. Table 8 further reports category-wise agreement for the final FS/SF/Col label.

D.2 Automatic Judge Analysis

We additionally evaluate whether off-the-shelf MLLM judges can replace human annotation. Each AI judge is given the same evaluation context as human annotators: the first-frame illusion image, the prompt without illusion description, and the generated video. The judge answers the same SA/HRC/VS-style questions and produces a final FS/SF/Col label. We then compare its predictions against human annotations.

Table 7: **Human annotation agreement.** We report pre-adjudication agreement between two independent annotators. The criteria and derived final labels are defined in Section 4.

Criterion	Agreement (%)	Cohen's κ
SA	95.9	0.767
HRC	92.5	0.846
VS	97.2	0.883
Final Label	90.4	0.838

Table 8: **Category-wise human annotation agreement.** Pre-adjudication agreement for final FS/SF/Col label across illusion categories.

Category	Agreement (%)	Disagreement (%)	Cohen's κ
Shadow	94.9	5.1	0.915
Refraction	93.5	6.5	0.864
Reflection	92.2	7.8	0.871
Perspective	90.2	9.8	0.829
Painting	88.9	11.1	0.819
Imaginative	86.1	13.9	0.669
Projection	85.7	14.3	0.778

As shown in Table 9, the AI judges show limited agreement with human annotations. Although Doubao-Seed2-lite (Seed Team, 2026) achieves higher average accuracy than Qwen3-VL-Flash (Alibaba Cloud, 2026), its Cohen's κ remains low under both reasoning and no-reasoning settings. This suggests that current MLLM judges are not reliable enough to replace human evaluation for VIVG-Bench. In particular, AI judges often struggle to distinguish Stable Failure from Full Success when a video appears visually coherent but still preserves the misleading first-frame interpretation. Therefore, all main results in VIVG-Bench are based on human annotations, while AI judging is used only as an auxiliary analysis.

Table 10: **Effect of removing trigger prompts.** Image-only removes target motion, so SA, FS, and SF are –; HRC† is a diagnostic score, not the main benchmark HRC label.

Setting	SA ↑	HRC† ↑	VS ↑	FS ↑	SF ↓	Col ↓
Prompt w./o. illusion description	98.56	65.81	96.47	65.01	31.46	3.53
Image only	–	46.51	93.18	–	–	6.82
Δ	–	**-19.30**	-3.29	–	–	+3.29

Table 11: **Effect of stating that a visual illusion is present.** The added sentence says that the scene contains a visual illusion. It does not name the category, the mechanism, or the expected outcome.

Setting	SA ↑	HRC ↑	VS ↑	FS ↑	SF ↓	Col ↓
Prompt w./o. illusion description	99.84	65.81	96.47	65.65	30.82	3.53
Prompt w. illusion cue	99.68	69.50	94.86	68.54	26.32	5.14
Δ	-0.16	**+3.69**	-1.61	**+2.89**	-4.50	+1.61

E Additional Prompt-side Analysis Details

E.1 Prompt w/o Illusion Description vs Image Only

Here we compare prompts without illusion descriptions with an image-only diagnostic variant on Grok Imagine. The image-only setting removes the prompt and provides only the illusion image as input. Because no target motion or interaction is specified, SA, FS, and SF are not applicable in this setting and are marked as – in Table 10. For this control only, HRC† measures whether spontaneous temporal changes expose relation-relevant evidence compatible with the image cues; it is a diagnostic score and is not used as the main benchmark HRC label.

This analysis is intended to test whether the prompt provides a dynamic condition that makes the relevant physical constraint testable. The observed drop in the diagnostic score indicates that the trigger is important for testing category-specific physical constraints. At the same time, VS stays high; without a specified motion, many videos never probe the constraint, so the diagnostic-score drop is not the same kind of violation.

Table 9: **AI judge comparison across three settings.** We compare three off-the-shelf AI judging settings against human annotations. Accuracy and Macro-F1 are reported as percentages; Cohen's κ is reported on its standard scale.

Setting	Avg Acc	Avg Macro-F1	Avg Cohen's κ
Qwen3-VL-Flash (Alibaba Cloud, 2026)	45.23	30.35	0.008
Doubao-Seed2-lite (reasoning=medium) (Seed Team, 2026)	58.00	40.48	0.179
Doubao-Seed2-lite (no reasoning) (Seed Team, 2026)	55.10	41.16	0.158

E.2 Prompt w/o Illusion Description vs an Illusion-Presence Cue

This control tests whether telling the model that a visual illusion is present improves constraint-consistent generation. It is run only on Grok Imagine and was recomputed on a re-selected paired subset, so the baseline in Table 11 is not the same returned-clip set as Table 1 or Table 10. Within Table 11, both rows use that subset. Each video is judged with the same SA, HRC, and VS protocol, and the aggregate results are reported in Table 11.

The baseline is the benchmark prompt without an illusion description: it specifies the intended motion or interaction and does not name the illusion. The enhanced prompt keeps that text unchanged and appends one sentence: "This scene contains a visual illusion." This sentence does not name the illusion category, does not describe the mechanism, and does not state the expected physical outcome. For example, it does not say that a painting is flat, that a forced-perspective pair is separated in depth, or how the depicted objects should move.

The comparison therefore asks whether awareness that an illusion is present is sufficient. It does not give the model the answer. On the paired subset, as reported in Section 6, HRC rises from 65.81% to 69.50% and FS from 65.65% to 68.54%, while Collapse rises from 3.53% to 5.14%. Noticing the illusion produces only a modest gain and does not reliably yield stable, constraint-consistent video.