Are Video Models Ready as Zero-Shot Reasoners?
An Empirical Study with the MME-CoF *Benchmark*

Supplementary Material

Overview

We organize this supplementary material as follows.
- Section A: Remaining deep-dive analysis on Veo-3.
- Section B: Related work.
- Section C: Additional implementation details.
- Section D: Statistics of MME-CoF.
- Section E: Limitations and future work.

A. Remaining Deep-Dive Analysis on Veo-3

Due to space limitations, we only provide task descriptions and partial qualitative analyses for tasks *1) - 5)* in the main paper Section 2. In this supplementary material, we additionally provide: *(i)* detailed definitions of the three-level performance criteria, *(ii)* the data sources, and *(iii)* additional representative examples for tasks *1) - 5)*. We also present the complete analysis for tasks *6) - 12)*, including task descriptions, evaluation criteria, and expanded qualitative results and analyses. For clarity, we restate the full list of tasks below:

1) Visual Detail Reasoning
2) Visual Trace Reasoning
3) Real-world Spatial Reasoning
4) 3D Geometry Reasoning
5) Physics-based Reasoning
6) 2D Geometry Reasoning
7) Rotation Reasoning
8) Table and Chart Reasoning
9) Object Counting Reasoning
10) GUI Reasoning
11) Embodied Reasoning
12) Medical Reasoning

A.1. Visual Detail Reasoning

In Section 2.2, we provide the task description and representative examples for the visual detail reasoning task. In this section, we further include the detailed definition of the three-level performance criteria and the data source used in this category.

Definition of *Good* / *Moderate* / *Bad*. We define the three-level evaluation criteria as follows:

✓ *Good*: The reasoning video accurately centers on the correct target region, clearly resolves the relevant attribute, such as color, texture or position, and maintains sharp, stable and natural rendering throughout the sequence. There are no visible frame drops, artifacts or unintended motion.

~ *Moderate*: The region of interest is approximately correct, and the attribute remains inferable, but the sequence suffers from minor blur, incomplete framing, slight instability mild unnatural motion, or sometimes deviates from the textual instruction and produces a plausible but unaligned or self-directed visual interpretation, limiting confident interpretation.

✗ *Bad*: The target region is incorrect or ambiguous, the attribute cannot be reliably inferred, or the video exhibits severe artifacts: abrupt frame jumps, major jitter, unintended zoom or crop, extraneous objects interfering, or conspicuous quality degradation that obstructs the reasoning task altogether.

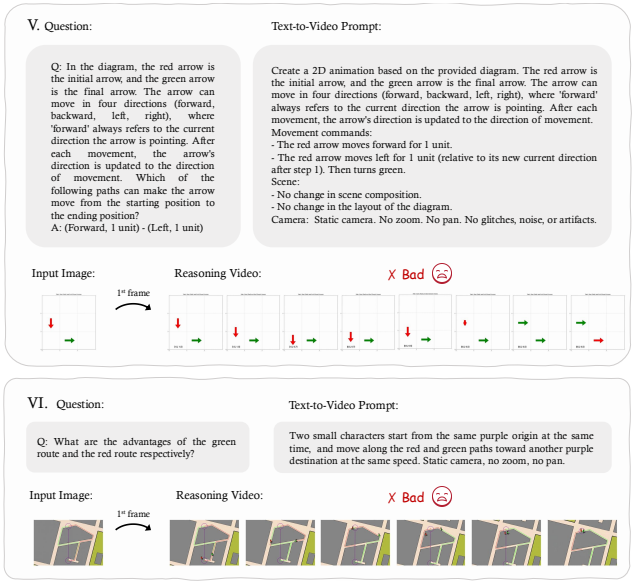

Figure 1. **Additional showcase of Visual Trace Reasoning by Veo-3.** The examples highlight long-horizon planning breakdowns, inconsistent arrow/trajectory rendering, and failures to preserve comparative or sequential information across frames.

Data Source. We sample data from the *V*Bench* [56], which provides a comprehensive set of evaluation dimensions including spatial relationship and color/attribute consistency tasks.

A.2. Visual Trace Reasoning

In Section 2.3, we describe the setup for the visual trace reasoning task and show several representative examples. Here, we provide the detailed three-level performance definition, the data sources, and additional qualitative cases.

Definition of *Good* / *Moderate* / *Bad*. We rate the performance according to the following criteria:

✓ *Good*: Each movement step is depicted continuously and logically toward the correct goal. The motion is smooth, temporally consistent, and follows causal order with no skipping, stuttering, or direction reversal.

~ *Moderate*: The overall trajectory roughly aligns with the intended sequence, but small discontinuities, timing irregularities, or partial missteps occur. The reasoning path remains interpretable, and the goal can still be inferred.

✗ *Bad*: Key steps are missing, reversed, or illogical. The sequence shows abrupt jumps, inconsistent object trajectories, or goal confusion, breaking the temporal and causal coherence of the reasoning process.

Data Source. We select samples from *MVoT* [29], *FrozenLake* [3, 57], *MiniBehavior* [24], *RBench-V* [17], *SpatialVizBench* [51], and *OmniSpatial* [21], which provide controlled multi-step environments for evaluating temporal reasoning, sequential planning, and causal continuity in visual simulations.

Additional Example and Analysis. As shown in Figure 1, case V reveals difficulty grounding abstract movement rules, producing inconsistent arrow trajectories. Case VI produces visually plausible motions along individual paths but fails to preserve or present the comparative information required for contrastive reasoning.

A.3. Real-World Spatial Reasoning

In Section 2.4, we introduce the real-world spatial reasoning task and present a subset of qualitative results. In this section, we include the full definition of the three-level evaluation criteria and specify the data sources used in this category.

Definition of *Good* / *Moderate* / *Bad*. We define the evaluation criteria in three levels:

✓ *Good*: Scene orientation, reference frame, and viewpoint are consistent and correctly represent spatial relations. The camera remains steady and the motion is natural.

~ *Moderate*: Scene roughly matches the instruction but contains small perspective errors, unnatural transitions, or partial mirroring. Motion remains interpretable but not physically coherent.

✗ *Bad*: Reference frame or direction is wrong; viewpoint shifts abruptly or inconsistently. Video suffers from strong camera drift, disorienting motion, or spatial chaos.

Data Source. To evaluate on orientation and layout reasoning, we specifically sample data from *MMSI-Bench* [59], *OmniSpatial* [21], *CoreCognition* [33] and *RefSpatial* [64]. Also, the tasks of perspective taking and spatial interaction are selected from the *OmniSpatial* dataset [21].

A.4. 3D Geometry Reasoning

In Section 2.5, we outline the 3D geometry reasoning task and discuss several illustrative examples. Here, we further detail the three-level performance criteria and summarize the data sources used for 3D geometry reasoning.

Definition of *Good* / *Moderate* / *Bad*. We categorize the model's performance into three levels:

✓ *Good*: Transformations like folding, rotation and assembly are geometrically correct, visually smooth, and continuous, maintaining structural integrity and realistic motion. No broken edges, jumps, or spatial artifacts.

~ *Moderate*: Transformations are partially correct but show local misalignment, unrealistic deformation, or discontinuous motion; geometry is roughly interpretable but imperfect.

✗ *Bad*: Transformation fails. For example, wrong fold, structure collapse, or impossible geometry. Motion is erratic, discontinuous, or visually implausible, breaking the sense of physical realism.

Data Source. To construct diverse and representative evaluation data, we adapt tasks from established geometric spatial reasoning datasets, including the *3D-Text-Instruct* and *Folding Nets* subsets of the *STARE* benchmark [31], the *BlockMoving* subset from the *SpatialViz-Bench* [51], as well as *VisuLogic* [58] and *OmniSpatial* [21].

A.5. Physics-based Reasoning

In Section 2.6, we provide the high-level description of the physics-based reasoning task together with representative qualitative cases. In this section, we additionally specify the three-level performance definition and the data sources for the physics-based reasoning category.

Definition of *Good* / *Moderate* / *Bad*. We rate the performance according to the following criteria:

✓ *Good*: The motion sequence adheres to physical laws such as gravity, momentum, and energy conservation. Object interactions are realistic and

temporally smooth, and the visual outcome remains coherent and credible throughout.

~ *Moderate*: The physical relations are approximately correct but include minor inconsistencies, such as irregular acceleration, timing mismatch, or slight violation of conservation. The overall motion remains interpretable and visually plausible.

✗ *Bad*: The motion is physically implausible or visually chaotic—objects float, stop abruptly, or behave contrary to basic causal principles. Severe artifacts or temporal discontinuities disrupt the perception of a coherent physical process.

Data Source. We draw samples from *MMMU* [60], *CoreCognition* [33], *ScienceQA* [37], related physical reasoning subsets of *RBench-V* [16], and *SpatialViz-Bench* [51], covering scenarios such as object collisions, pendulum motion, frictional sliding, and optical or magnetic interactions. Additional problems are collected from online resources [1] [2].

A.6. 2D Geometry Reasoning

In this section, we provide a complete analysis of Veo-3 [13] on the 2D geometry reasoning task.

Task Description and Evaluated Aspects. To assess a model's competence in 2D geometric reasoning, we evaluate its zero-shot performance on planar geometric construction tasks. These tasks involve drawing geometric relations by connecting points, adding auxiliary lines, and moving geometric shapes. The evaluation focuses on whether the generated constructions or movements accurately reflect the described geometric relationships and adhere to the given instructions, while maintaining smooth, stable operations that ensure visual clarity and coherence throughout the process.

Definition of *Good* / *Moderate* / *Bad*. We rate the performance according to the following criteria:

✓ *Good*: Constructions and movements are geometrically accurate and visually smooth. Endpoints, intersections, angles, and motion trajectories align correctly with the instructions. Both drawing and movement processes are stable, fluid, and natural, resembling human sketching or manipulation.

~ *Moderate*: Constructions and movements roughly follow the intended geometry but exhibit minor inaccuracies in line placement, shape alignment, trajectory, or smoothness. Some local jitter or abrupt motion may appear, but the overall structure and motion remain interpretable.

[1] physicstasks.eu
[2] physics.info

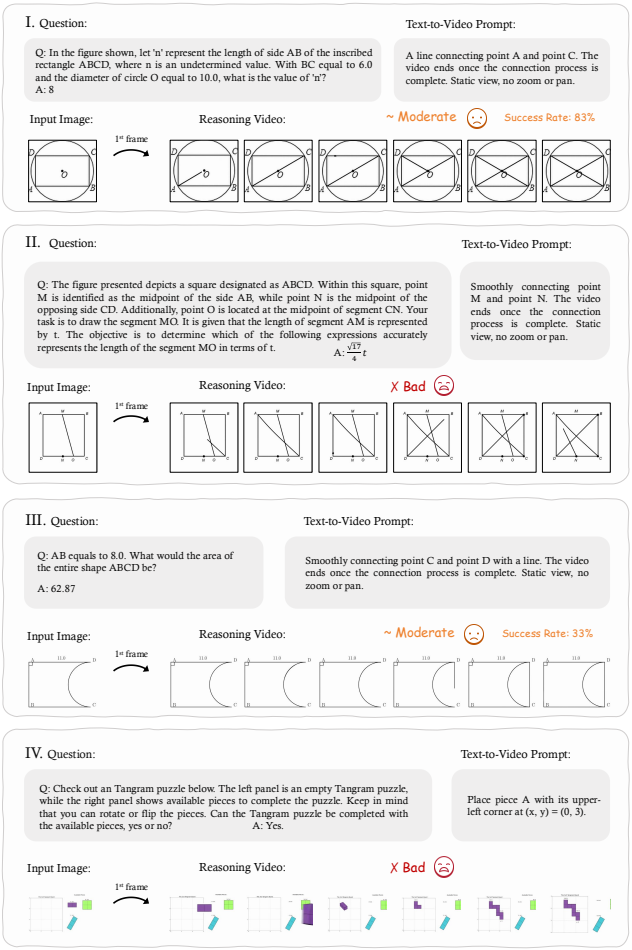

Figure 2. **Showcase of 2D Geometry Reasoning by Veo-3 (Part I).** While Veo-3 shows potential in recognizing simple patterns, it lacks the robust constraint awareness essential for accurate geometric manipulation.

✗ *Bad*: Constructions or movements deviate substantially from geometric correctness. Lines or shapes may be misplaced, disconnected, or moved in a chaotic or discontinuous manner (*e.g.* jittering, overlapping, or distorted paths), leading to visual instability and loss of interpretability.

Data Source. The evaluation data are drawn from multiple established sources, including the *Geo170k* dataset [11], the *VarsityTutors* subset of *Math-PUMA* [65] dataset, the *line-connection* subset of *RBench-V* [17], the *MAVIS-Gen* [61], *Tangram Puzzle* and *2D Text Instruct* subsets of the *STARE* [31] benchmark, and data from *VAT* [35].

Example and Analysis. The representative examples of the 2D geometry reasoning task are presented in Figures 2 and 3. Veo-3 demonstrates a foundational capability for

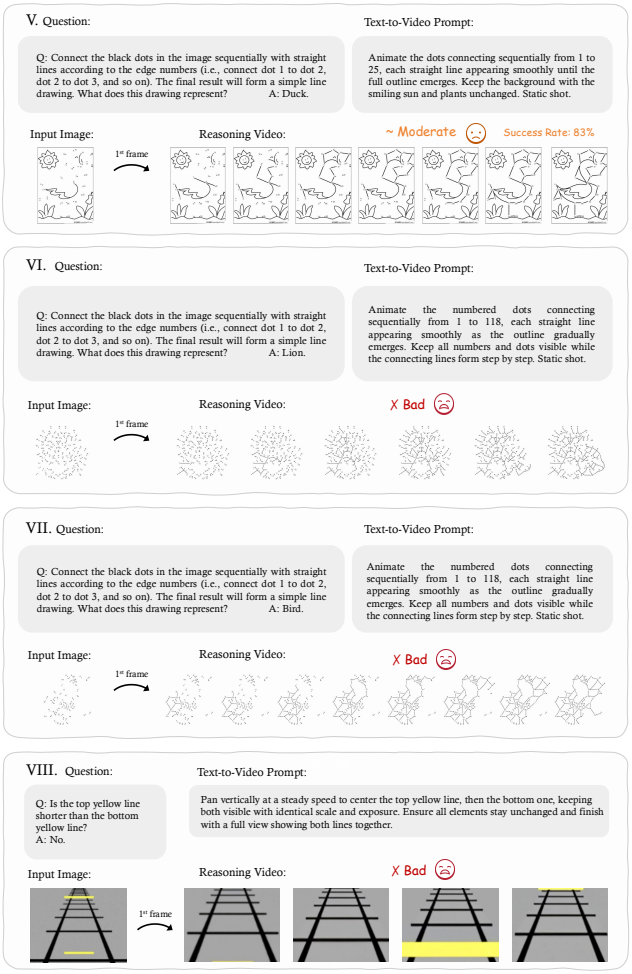

Figure 3. **Showcase of 2D Geometry Reasoning by Veo-3 (Part II).** Veo-3's reasoning abilities are further challenged by complex sequential instructions and the need to preserve structural integrity.

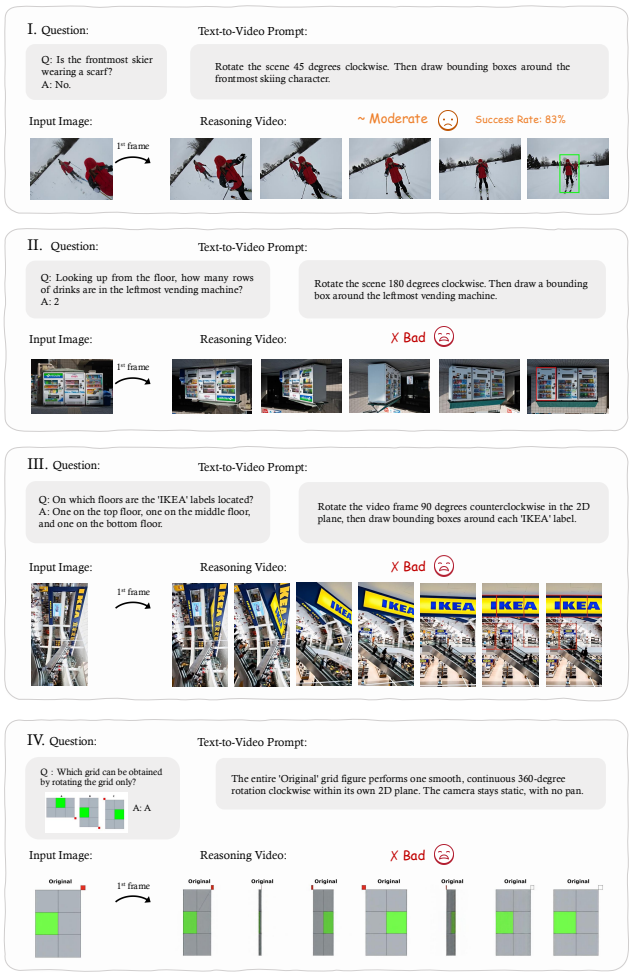

Figure 4. **Showcase of Rotation Reasoning by Veo-3.** Veo-3 struggles in complex scenes. However, its foundational grasp of simple rotations signals its potential to support rotation-based reasoning tasks.

simple geometric connection tasks, correctly identifying and linking elements in straightforward scenarios like in case III. However, this basic competence is inconsistent. The model often prioritizes producing visually symmetric or semantically meaningful patterns rather than strictly adhering to geometric instructions (cases I and II). Furthermore, case II reveals instances where the model unintentionally modifies the original figures, indicating a limited awareness of geometric constraints and poor spatial consistency. When tackling more complex connection tasks, the model frequently fails to interpret the intended drawing order or point indices, resulting in incorrect connection sequences, as demonstrated in cases V, VI, and VII. This is often coupled with an inability to control task termination, as the model tends to continue drawing beyond the required constructions. Finally, for tasks involving the movement of geometric shapes in cases IV and VIII, the model struggles to maintain geometric structural consistency throughout the motion.

A.7. Rotation Reasoning

In this section, we provide a complete analysis of Veo-3 [13] on the rotation reasoning task.

Task Description and Evaluated Aspects. The rotation reasoning task assesses the ability to reason about planar object rotation and maintain consistent spatial grounding under rotational transformations, thereby supporting subsequent reasoning processes. In each instance, the model is required to accurately rotate target objects within a fixed 2D plane while preserving the overall scene structure and structural consistency, followed by performing reasoning tasks like grounding and OCR. The evaluation focuses on both the accuracy of the rotation in terms of angle and direction, and

the precision of the resulting reasoning tasks.

Definition of *Good* / *Moderate* / *Bad*. Model outputs are categorized into three quality levels:

> ✓ *Good*: The rotation is accurate, complete, and strictly confined to the 2D plane, with no extraneous scene motion. The following reasoning tasks are completed correctly. Target objects remain precisely grounded after rotation.
>
> ~ *Moderate*: The rotation is largely correct but may be incomplete or slightly off-angle, though still confined to the 2D plane. The following reasoning tasks are mostly completed. Minor temporal or visual inconsistencies may appear, but do not alter the core 2D structure or object grounding.
>
> ✗ *Bad*: The model fails to perform the correct rotation, extends the transformation into 3D space, or introduces substantial scene distortion. Cannot complete the following reasoning task. The original 2D structure is altered, leading to inaccurate grounding of the target objects.

Data Source. To specifically assess the rotation reasoning task, we recruit some PhD-level experts with deep expertise in text-image reasoning to design the evaluation data manually, followed by the necessary review process. Each question is designed following the principle that it must involve a 2D rotation to reach the correct solution, ensuring the task genuinely probes rotational understanding rather than simple visual matching. Moreover, we sample data from the *2DRotation* subset from the *SpatialViz-Bench* [51], and reformulate the question into instructions for the video models.

Example and Analysis. The results are shown in Figure 4. In case I, we find that Veo-3 handles small-angle rotations and simple planar scenes reasonably well, demonstrating a basic grasp of rotational motion. However, in more complex scenarios like cases II, III, and IV, the model often ignores the 2D rotation constraint and inadvertently alters the 3D structure, resulting in incorrect rotations and degraded spatial grounding. Such errors frequently propagate to downstream tasks, such as OCR in case III, or object localization in case II, due to inconsistencies in post-rotation alignment. These observations suggest that the reasoning behavior of Veo-3 remains more pattern-driven rather than principle-driven. However, as it demonstrates a partial understanding of planar rotation, this can to some extent facilitate subsequent reasoning tasks.

A.8. Table and Chart Reasoning

In this section, we provide a complete analysis of Veo-3 [13] on the table and chart reasoning task.

Task Description and Evaluated Aspects. The table and chart reasoning task requires the model to identify and focus on the key elements within visualizations or tabular data. For evaluation, we further consider how effectively the model identifies the regions relevant to the query and whether it can transition smoothly and visually coherently to these areas, preserving clarity, continuity, and proper scaling.

Definition of *Good* / *Moderate* / *Bad*. We rate the performance according to the following criteria:

> ✓ *Good*: Camera precisely focuses on the correct chart or table segment, smoothly highlighting or zooming into the queried data (*e.g.* correct year, category, or value). Motion is continuous, the chart and table remain clear, and no distortion or overexposure occurs.
>
> ~ *Moderate*: Camera approximately focuses on the right region but partially misses boundaries, introduces slight blur, or transitions abruptly. Data can still be inferred.
>
> ✗ *Bad*: Video fails to locate the correct region or changes the chart or table geometry unnaturally. Motion jitter, scaling errors, or artifacts make data unreadable or misleading.

Data Source. We use samples from the *ChartQA* [40] dataset and *TableVQA-Bench* [26].

Example and Analysis. For charts, as presented in cases I, II and III in Figure 5, Veo-3 can often zoom into an approximately correct region but lacks the precision needed to accurately locate the queried data. For tables, as shown in case IV, Veo-3 fails to correctly identify the required element and tends to select entries randomly. The model also frequently adds, modifies, or distorts existing chart and table elements, resulting in visual inconsistencies that undermine the accuracy of chart interpretation.

A.9. Object Counting Reasoning

In this section, we provide a complete analysis of Veo-3 [13] on the object counting reasoning task.

Task Description and Evaluated Aspects. In this category, we focus on the ability to accurately enumerate objects within a 2D or 3D scene. In each instance, the model is

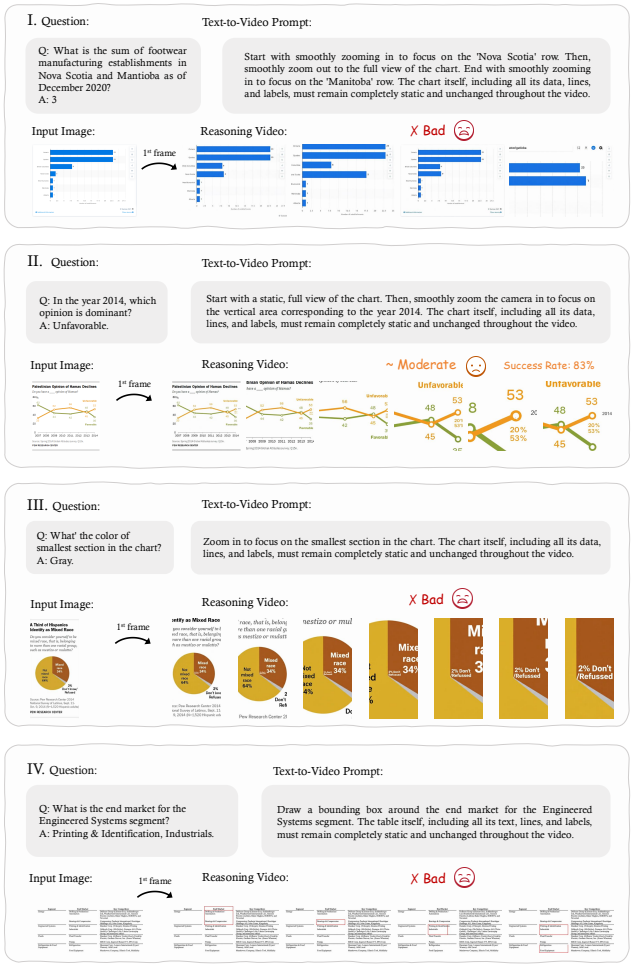

Figure 5. **Showcase of Table and Chart Reasoning by Veo-3.** Veo-3 demonstrates an initial ability to focus on relevant data regions but lacks the precision and consistency required for reliable visual analysis.

required to identify, ground, and count target objects, typically by highlighting, drawing bounding boxes, applying numerical labels, or panning. The evaluation focuses on the accuracy of the count and the precision of the spatial grounding, performed within a scene that remains static or experiences only minimal motion, ensuring the counting process is not influenced.

Definition of *Good* / *Moderate* / *Bad*. Model outputs are categorized into three quality levels:

> ✓ *Good*: The model precisely highlights, draws bounding boxes around, or labels the objects with correct numbers, and performs smooth and controlled panning when necessary to cover all targets. Motion is continuous, and the scene remains static or experiences only slight changes that do not influence the counting process.
>
> ~ *Moderate*: The model approximately highlights or draws bounding boxes around the objects, or performs panning with minor instability or incomplete coverage. Objects or the scene may move or change slightly, but this does not strongly affect the counting process.
>
> ✗ *Bad*: The model fails to correctly highlight, label, or draw bounding boxes around the objects, or pans erratically such that parts of the scene are missed or revisited unnecessarily. Objects or the scene move or change substantially, severely affecting the counting process.

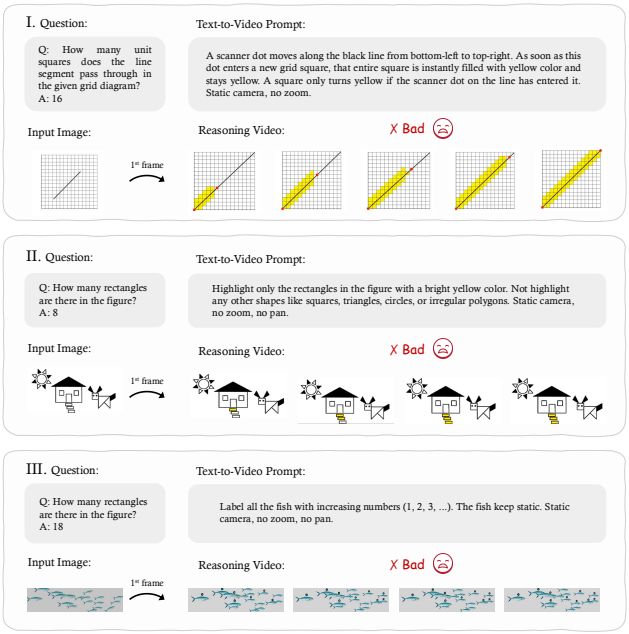

Figure 6. **Showcase of 2D Object Counting Reasoning by Veo-3.** Veo-3's lack of spatial control often introduces object motion, undermining the stability and accuracy of the counting process.

Data Source. The 2D object counting data are sampled from the *counting* subset of *RBench-V* [17]. The 3D object counting data are from the *Super-CLEVER* dataset [34], *CoreCognition* [33] and *VAT* [35].

Example and Analysis. The results are shown in Figures 6 and 7. In the 2D counting tasks from cases I to III, objects frequently move or change during the process, negatively impacting counting stability and accuracy. In the 3D counting tasks, Veo-3 successfully handles simple grounding and counting scenarios, as demonstrated in case V, but struggles with scenes involving complex materials or geometric variations in cases VI and VII, leading to inaccurate counts.

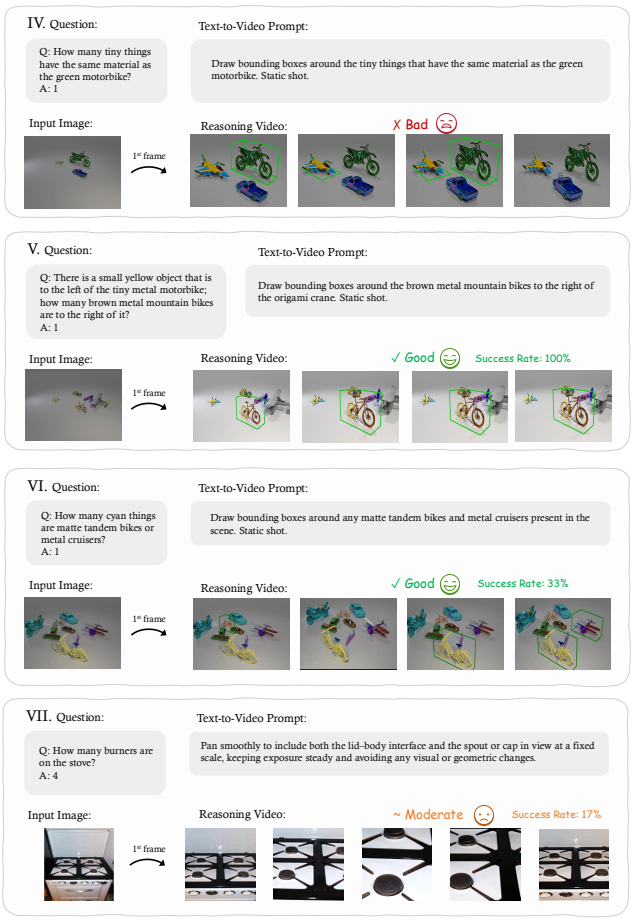

Figure 7. **Showcase of 3D Object Counting Reasoning by Veo-3.** Veo-3's basic 3D counting abilities are challenged by complex materials, geometric variations, and imprecise camera control.

Additionally, in the panning process of case VII, the camera fails to precisely move to the regions containing all target objects, further hindering the counting process.

A.10. GUI Reasoning

In this section, we provide a complete analysis of Veo-3 [13] on the GUI reasoning task.

Task Description and Evaluated Aspects. In the Graphical User Interface (GUI) reasoning task, we focus on the capability to understand and interact with graphical user interfaces across different operating systems, including Android, Linux, and Web environments. In each instance, the model is required to perform actions, such as clicking on specific UI elements. The evaluation focuses on the accuracy of the click and the temporal coherence of the interaction, ensuring the scene and irrelevant UI elements remain consistent.

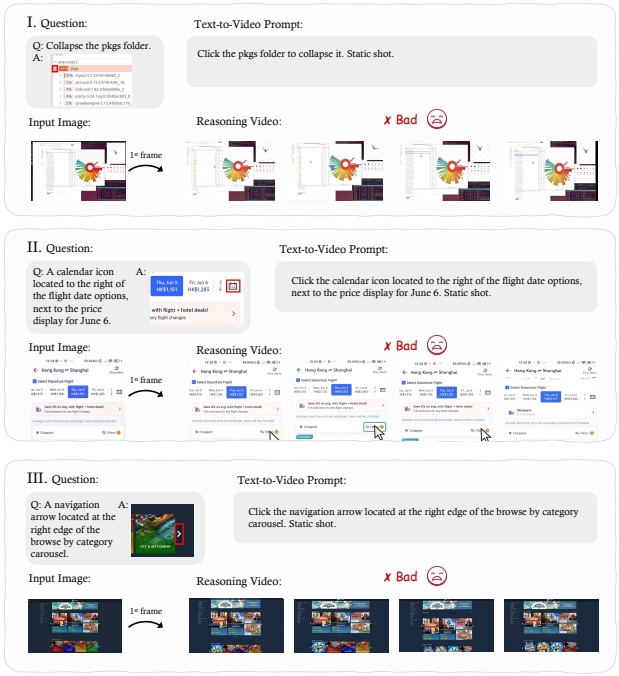

Figure 8. **Showcase of GUI Reasoning by Veo-3.** Veo-3's attempts at graphical interface interaction exhibit visual inconsistencies and logical inaccuracies, indicating only a shallow grasp of underlying GUI logic. Note that the answer to each question is a bounding box. For visual clarity, screenshots with the ground-truth bounding boxes are shown.

Definition of *Good* / *Moderate* / *Bad*. We define the evaluation criteria in three levels:

> ✓ *Good*: The click is precise, with no extraneous actions. No superfluous icons appear, and the original data and icons remain unchanged.
>
> ~ *Moderate*: The click is precise but may be accompanied by minor extraneous actions. Superfluous icons might appear but do not obscure the click target, and original data or icons show only slight alterations.
>
> ✗ *Bad*: The click is imprecise or erratic. Original data and icons are significantly altered, hindering judgment and assessment.

Data Source. We select Android, Mac, IOS, Web and Linux data from *MMBench-GUI* [52], *ScreenSpot-Pro* [30] and *OmniSpatial* [21].

Example and Analysis. Across the three cases in Figure 8, Veo-3 fails to accurately capture the correct click position and often exhibits inconsistencies between the click location and the resulting on-screen effect. In addition, it occasionally

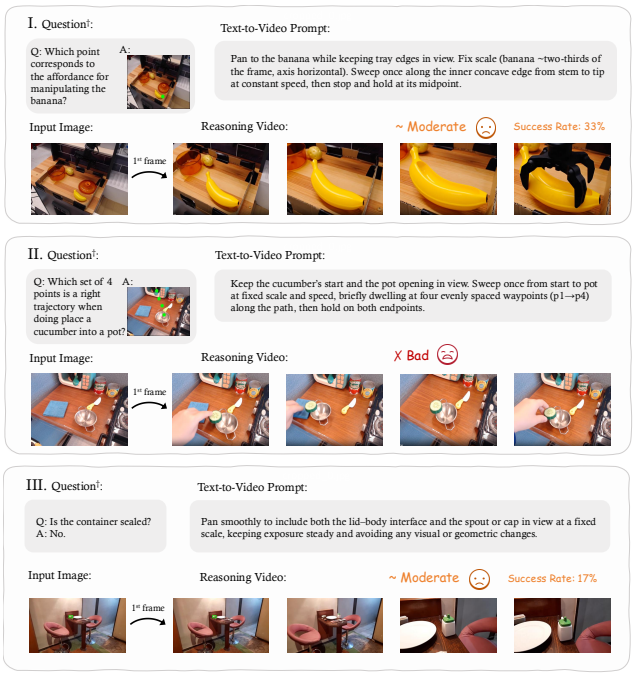

Figure 9. **Showcase of Embodied Reasoning by Veo-3.** It illustrates plausible static affordance detection in simple settings, common workaround/hallucination behaviors for dynamic manipulations, and failures to reliably localize or preserve manipulation-relevant context. † Green points in the answer image denote ground-truth points or trajectories.

alters or generates new icons and text, which can interfere with judgment. In the Web system in case III, however, the model demonstrates partial GUI responsiveness and provides some degree of visual feedback.

A.11. Embodied Reasoning

In this section, we provide a complete analysis of Veo-3 [13] on the embodied reasoning task.

Task Description and Evaluated Aspects. This category evaluates the model's potential to perceive and reason about object affordances and manipulation dynamics. It involves recognizing both static and dynamic affordances, as well as identifying manipulation-relevant object and scene attributes. Evaluation focuses on two aspects: *(i)* the generation of stable and contextually relevant visual sequences, and *(ii)* the maintenance of reasoning fidelity without resorting to implausible planning shortcuts or hallucinated interactions.

Definition of *Good* / *Moderate* / *Bad*. We define the evaluation criteria in three levels:

✓ *Good*: The sweep/framing covers all candidates fairly (equal or near-equal dwell), centers the manipulation-relevant geometry (*e.g.* handle + frame/gap, lid-body interface, hinge side) with crisp focus and stable scale; no cropping of key context; no content alterations.

~ *Moderate*: The view roughly includes the right region(s) but with minor bias or coverage issues: slight off-center, brief under-exposure of one candidate, small motion jitter, or shallow context (still enough to infer).

✗ *Bad*: The camera misses or biases the evidence (*e.g.* lingers only on one point, crops away the hinge/rail, over-zooms a non-relevant patch), introduces distortion/content edits, or produces footage from which a fair decision cannot be made.

Data Source. We select samples from *Robobench* [39] for the analysis. In addition to a general understanding of static attributes, we also sample data to assess whether Veo-3 can perform direct reasoning on tasks involving the generation of static and dynamic affordances.

Example and Analysis. As shown in Figure 9, Veo-3 demonstrates the ability to comprehend objects within real-world scenes. However, its capacity for assisting visual reasoning in embodied scenarios remains constrained by insufficient stability. As illustrated in case I, when provided with a clearly defined object for manipulation, Veo-3 is capable of generating plausible manipulation affordances. When it comes to dynamic affordances, Veo-3 tends to employ workarounds to compensate for its planning deficiencies, as evidenced in case II, where it generated a new cucumber instead of the intended object. With respect to static attributes, Veo-3 struggles to accurately differentiate visual prompts and misidentifies the position of containers. As shown in case III, the green box, intended to specify the location of the container, inadvertently led Veo-3 to produce hallucinations.

A.12. Medical Reasoning

In this section, we provide a complete analysis of Veo-3 [13] on the medical reasoning task.

Task Description and Evaluated Aspects. This category assesses the model's ability to localize lesions or structures, identify relevant attributes (*e.g.* side, lobe), recognize pathological patterns (*e.g.* "jump distribution"), and make binary decisions (*e.g.* presence or absence). The evaluation focuses on both the correctness of object manipulation and the visual stability of the surrounding regions.

Definition of *Good* / *Moderate* / *Bad*. We define the evaluation criteria in three levels:

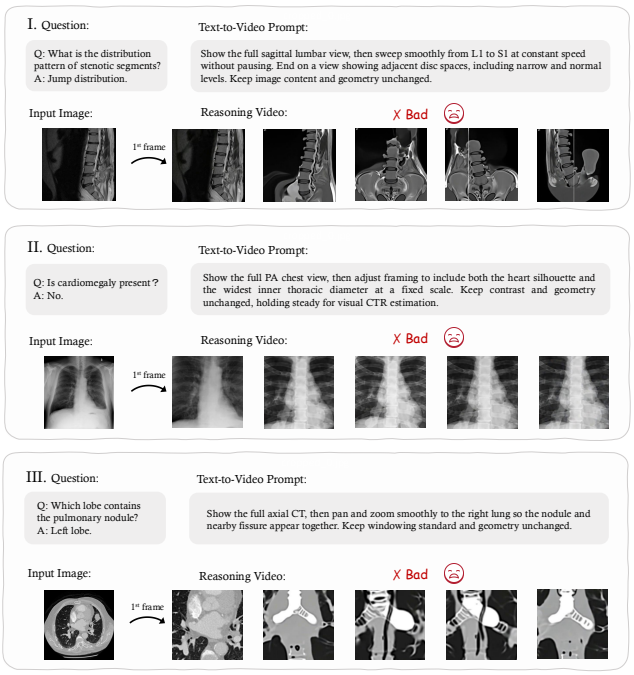

Figure 10. **Showcase of Medical Reasoning by Veo-3.** As shown in cases I and III, Veo-3 fails to maintain the shape of the rest of medical organization. Veo-3 also can not understand and precisely locate the mentioned medical terminology in the prompt, as demonstrated in case II.

✓ *Good*: The camera cleanly settles on the correct anatomical level/lesion, with clear margins and readable context; motion is reasonable; no geometric distortion or content alteration.

~ *Moderate*: The view roughly covers the right area but is slightly off (partial coverage, mild blur, small framing mistakes). The general shape of the tissue or organ can still be observed.

✗ *Bad*: The video misses the target region or introduces distortions/crops that hide key cues. Tissues or organs begin to distort. Misleading results due to confusion of medical terminology.

Data Source. We select samples representing different body parts from the *ViTAR* [5], *PMC-VQA* [62] and *Med-VQA* [10] dataset.

Example and Analysis. We showcase the evaluation results in Figure 10. Veo-3 retains the ability to manipulate images when dealing with medical images. However, due to its lack of medical knowledge, Veo-3 struggles to accurately manipulate the correct objects when instructions include medical terminology. This phenomenon is evident across all cases. Furthermore, Veo-3 cannot model medical organs effectively. When performing operations such as zooming in, the medical images suffer from significant distortion, resulting in a substantial loss of detail.

B. Related Work

Video Models. Video models have been progressively evolving both in the fields of video understanding and generation. For video understanding methods, earlier approaches, such as MViT [7], Video Swin Transformer [36], and VideoMAE [48], aim to learn a robust representation that fosters downstream tasks. With the rise of LLMs, recent approaches encode videos as tokens and exploit the language backbone for captioning [47], event localization [45], and high-level reasoning [20, 63]. Video generation models have also attracted much attention. Closed system, including OpenAI's Sora [4, 43], Runway's Gen-3 [44], Luma AI [38], and Google DeepMind's Veo series [12, 13], have exhibited impressive results. However, they remain inaccessible due to their closed-source nature. Open-source alternatives have recently become available: Stable Video Diffusion [2] introduces efficient training strategies, Hunyan-Video [27] proposes systematic scaling, and Wan-2.1 [49] presents an efficient 3D VAE with expanded pipelines.

Reasoning with Video. The advent of large reasoning models [14, 18, 19, 22, 23, 46, 54], such as OpenAI o1 [42] and DeepSeek-R1 [15], has spurred the development of video reasoning benchmarks. Most current methods [8, 32, 41] employ MLLMs specialized in video reasoning understanding. For example, Video-R1 [8] specifically targets temporal reasoning capabilities by introducing a temporal group relative policy optimization (GRPO) loss. VideoChat-R1 [32] focuses on spatio-temporal reasoning abilities by training with GRPO and rule-based rewards. A two-stage training strategy, combining SFT and RL, is used by VideoRFT [50]. When trained on vast collections of images and videos, this strategy boosts the model's ability to handle QA tasks, whether in general contexts or reasoning-focused ones. These methods primarily focus on enhancing specific types of question-answering or captioning tasks. Concurrently, [55] demonstrates the large potential of video generative models in video reasoning. These models have implicitly acquired world knowledge throughdemonstrates impressive performance on various tasks, includinging and reasoning capability. Yet, this direction has rarely been explored and only experimented with in zero-shot settings.

Evaluation of Video Models as Zero Shot Learner. Recently, several works have been exploring the zero-shot capability of video generation models in various domains, including general-purpose vision understanding [9, 55], medical imaging [28], and world models [53]. [55] conducts experi-

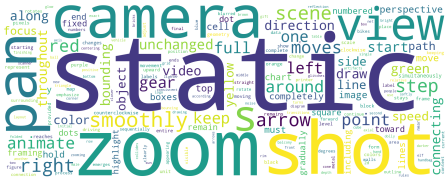

Figure 11. **Word Cloud of MME-CoF Benchmark.** It presents the distribution of key task concepts throughout the benchmark.

ments on Veo 3 with a variety of vision tasks that have not been explicitly included during training. The video model showcases surprising performance on multiple tasks like object segmentation, image editing, and even maze solving. [28] later adopts a similar paradigm to medical images understanding tasks and finds video generation models also show powerful capabilities, *e.g.* delineation of anatomical structures in CT scans, medical image segmentation, and even forecasting of future 3D CT phases. Besides, [53] shows that video generation models could also understand complex temporal causality and world knowledge in the real world, thereby serving as a world model [1, 25].

C. Additional Implementation Details

Prompt Design Style for MME-CoF. As mentioned in the main paper, to ensure consistency and fairness, all the prompts in the MME-CoF benchmark follow a unified style emphasizing explicit visual constraints, controlled motion, and minimal linguistic ambiguity. Prompts are encouraged to be written in imperative form and designed to reduce variance from language interpretation, focusing the model's behavior on the intended visual reasoning objective. The overall prompt design principles are as follows:

1) *Static camera and fixed viewpoint, unless motion is explicitly required by the task.*
2) *Stable spatial composition, consistent framing, and unchanging scene layout across frames.*
3) *Clear specification of allowed and disallowed changes (*e.g. *"no zoom, no pan, no dolly") to constrain camera dynamics.*
4) *Explicit temporal phrasing to control the pace of motion, using cues such as "instantly", "smoothly", or "step-by-step".*
5) *Avoidance of direct textual hints toward the answer; instructions are purely visual and task-oriented.*
6) *Inclusion of realistic phrasing and scene context to align with the model's natural video priors while minimizing artifacts.*

The standardized prompt style ensures that differences in output primarily reflect the model's internal reasoning potential rather than prompt variability.

Table 1. **Key Statistics of MME-CoF.**

Statistic	Number
Total entries	120
Total categories	12
Max prompt length	124
Avg prompt length	34.1

Additional Evaluation Protocol Details. To maintain consistency across all video samples, input images are padded to a 16:9 aspect ratio with a white background prior to generation. For quantitative evaluation, we uniformly sample 16 frames from each generated video and prompt Gemini-2.5-Pro [6] with a unified instruction, aligning with the evaluation criteria described earlier:

Evaluate whether the following video faithfully and coherently visualizes the process described in the text.

Video frames are provided as images sequentially.

Process description is: {DESCRIPTION}

Judge the video only based on visible evidence.
Rate each of the following five aspects on a 0–4 scale (0 = poor, 4 = excellent):

```
{
  "instruction_alignment": 0,    // How
      well the video follows the
      described structure and sequence
  "temporal_consistency": 0,     //
      Smoothness and continuity between
       frames
  "visual_stability": 0,         //
      Stability of appearance, motion,
      and viewpoint
  "content_fidelity": 0,         //
      Preservation of key elements
      without hallucination or loss
  "focus_relevance": 0           //
      Whether visual attention stays on
       the intended objects or regions
}
```

D. Statistics of MME-CoF

The MME-CoF benchmark comprises of 120 entries, across 12 different tasks. More details of the statistics are in Table 1. Also, we provide the word cloud of the benchmark in Figure 11.

E. Limitations and Future Work

Although MME-CoF covers many reasoning scenarios tailored for video models, several important dimensions remain underexplored (no benchmark can be truly exhaustive). For example, capabilities involving text rendering, ARC-style abstract reasoning, chess or board-game understanding, and rule-based gameplay are not yet included. Extending the benchmark to systematically evaluate these challenging domains would further enhance its comprehensiveness and diagnostic power.

In addition, many closed-source video models (e.g., Veo, Sora) internally employ LLM/LMM-based prompt rewriters before the actual generation process. Since these rewriters are inaccessible and the entire pipeline is treated as a black box, it becomes difficult to isolate their contribution from the core video generation model. Understanding how to factor out, or explicitly account for, the influence of such prompt rewriting modules is therefore essential. This also raises a broader conceptual question: what exactly should be considered "the video model"? Should the definition include the prompt rewriter as an integral component, or should it be evaluated separately? This remains an open problem and an important direction for future work.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 10

[2] 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. 9

[3] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016. 2

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 9

[5] Kaitao Chen, Shaohao Rui, Yankai Jiang, Jiamin Wu, Qihao Zheng, Chunfeng Song, Xiaosong Wang, Mu Zhou, and Mianxin Liu. Think twice to see more: Iterative visual reasoning in medical vlms. *arXiv preprint arXiv:2510.10052*, 2025. 9

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 10

[7] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6824–6835, 2021. 9

[8] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 9

[9] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *CVPR 2025 Highlight*, 2024. 9

[10] Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. Medthink: Explaining medical visual question answering via multimodal decision-making rationale. *arXiv preprint arXiv:2404.12372*, 2024. 9

[11] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023. 3

[12] Google DeepMind. Veo 2, 2024. Accessed: 2024. 9

[13] Google DeepMind. Veo-3 technical report. Technical report, Google DeepMind, 2025. 3, 4, 5, 7, 8, 9

[14] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 9

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 9

[16] Meng-Hao Guo, Xuanyu Chu, Qianrui Yang, Zhe-Han Mo, Yiqing Shen, Pei-Lin Li, Xinjie Lin, Jinnian Zhang, Xin-Sheng Chen, Yi Zhang, Kiyohiro Nakayama, Zhengyang Geng, Houwen Peng, Han Hu, and Shi-Min Hu. Rbench-v: A primary assessment for visual reasoning models with multi-modal outputs. 2025. 3

[17] Meng-Hao Guo, Xuanyu Chu, Qianrui Yang, Zhe-Han Mo, Yiqing Shen, Pei-lin Li, Xinjie Lin, Jinnian Zhang, Xin-Sheng Chen, Yi Zhang, et al. Rbench-v: A primary assessment for visual reasoning models with multi-modal outputs. *arXiv preprint arXiv:2505.16770*, 2025. 2, 3, 6

[18] Ziyu Guo, Renrui Zhang, Hongyu Li, Manyuan Zhang, Xinyan Chen, Sifan Wang, Yan Feng, Peng Pei, and Pheng-Ann Heng. Thinking-while-generating: Interleaving textual reasoning throughout visual generation. *arXiv preprint arXiv:2511.16671*, 2025. 9

[19] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, et al. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025. 9

[20] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline pro-

fessional videos. *arXiv preprint arXiv:2501.13826*, 2025. 9

[21] Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025. 2, 7

[22] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025. 9

[23] Dongzhi Jiang, Renrui Zhang, Haodong Li, Zhuofan Zong, Ziyu Guo, Jun He, Claire Guo, Junyan Ye, Rongyao Fang, Weijia Li, et al. Draco: Draft as cot for text-to-image preview and rare concept generation. *arXiv preprint arXiv:2512.05112*, 2025. 9

[24] Emily Jin, Jiaheng Hu, Zhuoyi Huang, Ruohan Zhang, Jiajun Wu, Li Fei-Fei, and Roberto Martín-Martín. Mini-behavior: A procedurally generated benchmark for long-horizon decision-making in embodied ai. *arXiv preprint arXiv:2310.01824*, 2023. 2

[25] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Huang Gao, and Jiashi Feng. How far is video generation from world model? – a physical law perspective. *arXiv preprint arXiv:2406.16860*, 2024. 10

[26] Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*, 2024. 5

[27] 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. 9

[28] Yuxiang Lai, Jike Zhong, Ming Li, Yuheng Li, and Xiaofeng Yang. Are video models emerging as zero-shot learners and reasoners in medical imaging? *arXiv preprint arXiv:2510.10254*, 2025. 9, 10

[29] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025. 2

[30] Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. *arXiv preprint arXiv:2504.07981*, 2025. 7

[31] Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. *arXiv preprint arXiv:2506.04633*, 2025. 2, 3

[32] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. 9

[33] Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, et al. Core knowledge deficits in multi-modal language models. *arXiv preprint arXiv:2410.10855*, 2024. 2, 3, 6

[34] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973, 2023. 6

[35] Dairu Liu, Ziyue Wang, Minyuan Ruan, Fuwen Luo, Chi Chen, Peng Li, and Yang Liu. Visual abstract thinking empowers multimodal reasoning. *arXiv preprint arXiv:2505.20164*, 2025. 3, 6

[36] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, 2022. 9

[37] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 3

[38] LumaLabs. Dream machine, 2024. Accessed: 2024. 9

[39] Yulin Luo, Chun-Kai Fan, Menghang Dong, Jiayu Shi, Mengdi Zhao, Bo-Wen Zhang, Cheng Chi, Jiaming Liu, Gaole Dai, Rongyu Zhang, Ruichuan An, Kun Wu, Zhengping Che, Shaoxuan Xie, Guocai Yao, Zhongxia Zhao, Pengwei Wang, Guang Liu, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. Robobench: A comprehensive evaluation benchmark for multimodal large language models as embodied brain, 2025. 8

[40] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. 5

[41] Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and Zhuochen Wang. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025. 9

[42] OpenAI. Openai o1 system card. https://openai.com/index/openai-o1-system-card/, 2024. Accessed: 2024-12-05. 9

[43] OpenAI. Sora 2 system card. Technical report, OpenAI, 2025. 9

[44] GA RunwayML. Introducing gen-3 alpha: a new frontier for video generation, 2024. 9

[45] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. 9

[46] Chengzhuo Tong*, Ziyu Guo*, Renrui Zhang*, Wenyu Shan*, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. Delving into rl for image generation with cot: A study on dpo vs. grpo. *arXiv preprint arXiv:2505.17017*, 2025. 9

[47] Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. G-veval: A versatile metric for evaluating image and video captions using gpt-4o. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7419–7427, 2025. 9

[48] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 9

[49] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 9

[50] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025. 9

[51] Siting Wang, Minnan Pei, Luoyang Sun, Cheng Deng, Kun Shao, Zheng Tian, Haifeng Zhang, and Jun Wang. Spatialviz-bench: An mllm benchmark for spatial visualization. *arXiv preprint arXiv:2507.07610*, 2025. 2, 3, 5

[52] Xuehui Wang, Zhenyu Wu, JingJing Xie, Zichen Ding, Bowen Yang, Zehao Li, Zhaoyang Liu, Qingyun Li, Xuan Dong, Zhe Chen, et al. Mmbench-gui: Hierarchical multi-platform evaluation framework for gui agents. *arXiv preprint arXiv:2507.19478*, 2025. 7

[53] Zeqing Wang, Xinyu Wei, Bairui Li, Zhen Guo, Jinrui Zhang, Hongyang Wei, Keze Wang, and Lei Zhang. Videoverse: How far is your t2v generator from a world model? *arXiv preprint arXiv:2510.08398*, 2025. 9, 10

[54] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 9

[55] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. 9

[56] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 1

[57] Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. Vsp: Assessing the dual challenges of perception and reasoning in spatial planning tasks for vlms. *arXiv preprint arXiv:2407.01863*, 2024. 2

[58] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025. 2

[59] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. 2

[60] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 3

[61] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv e-prints*, pages arXiv–2407, 2024. 3

[62] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 9

[63] Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489, 2025. 9

[64] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025. 2

[65] Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 26183–26191, 2025. 3